医療、政府・自治体、大学による
エコシステムの実証

論文集

目次

オープンデータ

ビジネス活用

プラットフォーム

経済分析

教育

マーケティング管理

リスク管理

医薬品開発

その他関連分野

統計解析

SASシステム

医薬品開発、SASシステム

SASユーザー会活動の紹介

金融／経済／システム

基調講演

チュートリアル

Let's データ分析

オープンデータ

生物多様性を探るために

～統計解析からわかった トンボと生物多様性について～

宇久村 三世
北海道札幌旭丘高等学校 生物部

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

5年間のトンボ相の解析から、トンボ相の多様度指数の変動と、正の相関、負の相関のある種を特定することができた。和文標題の論文名を言語解析したところ、トンボや生物多様性の研究の必要性がわかった。

キーワード: 生物多様性、ノシメトンボ、アジアイトトンボ

解析理由

- データの解析方法をマスターしたいと思ったから
- 自然の仕組みを知り、見た目にはわからない
  生態系の仕組みを客観的に表したいから
- 生物多様性は重要
  → 大量のデータを集めて分析し、
  研究を進めていく必要がある

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

トンボにした理由

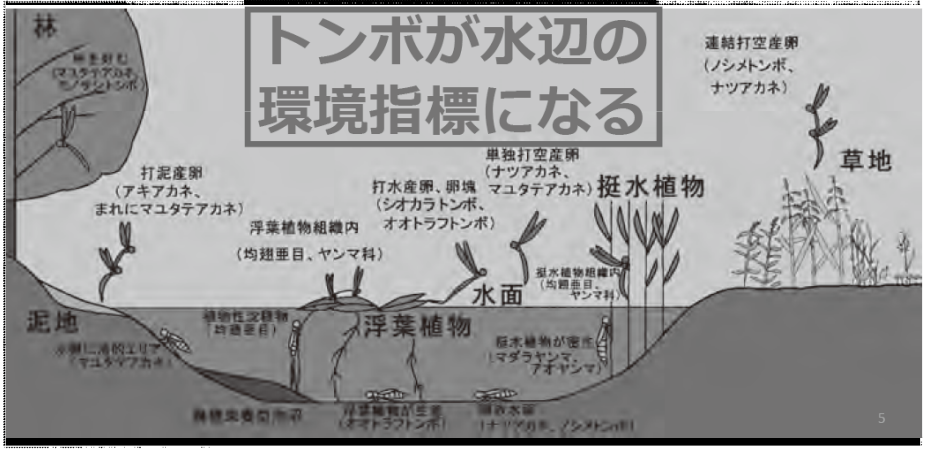
- 「勝ち虫」 → 兜や刀の鍔に
- 豊作をもたらす → 銅鐸に
- ことわざ、童謡の題材に
- 不退転の決意を表す

日本人にとって身近な生物

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

トンボにした理由

- 一生のうちに幅広く多様な環境を利用
- 産卵場所、ヤゴや成虫の生息場所も多様



トンボが水辺の環境指標になる

5

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

所有のデータ（トンボ相の定量調査の結果）

- 本校生物部が行ってきた、2009～2013年のトンボ相の定量調査の結果である
- 5～9月に月3回(年15回)の調査を5年間行った
- 石狩川下流域の沼や雨水調節池の5地点で行った
- それぞれの地点で造成後の経過年数が異なる
  → 一度に十数年の変動がわかる
- 5年間で、28種、13,584個体のトンボを採集した

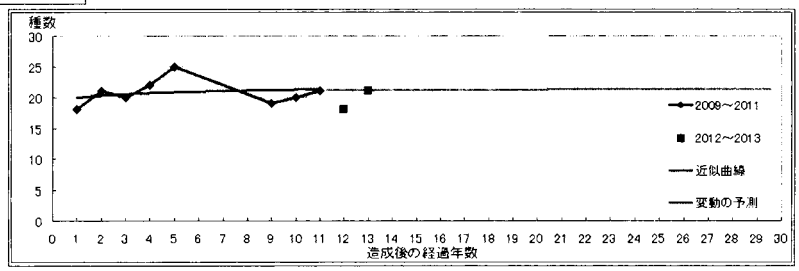
6

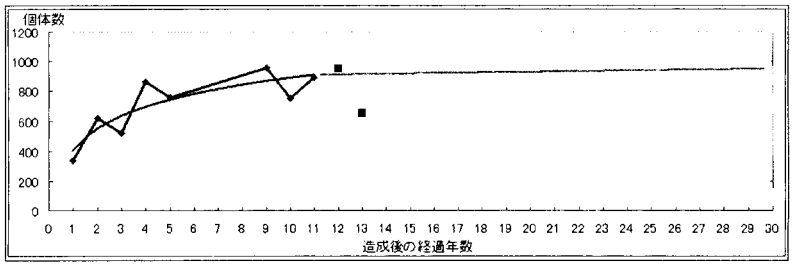
医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

種名・科名	調査地点、調査年(その地点ができてから何年経過しているか)																合計
	A地点	B地点		C地点			D地点					E地点			トンネウス沼		
	2011 1年目	2011 2年目	2012 3年目	2009 3年目	2010 4年目	2011 5年目	2009 9年目	2010 10年目	2011 11年目	2012 12年目	2013 13年目	2009 64年目	2010 65年目	2011 66年目	2012 40年目	2013 41年目	
アオイトトンボ科	54	78	89	100	133	191	184	125	136	135	109	54	90	35	252	164	1929
1 アオイトトンボ	31	50	31	56	79	137	112	82	58	50	44	5	7	2	21	47	812
2 オツネントンボ	23	28	58	44	54	54	72	43	78	85	65	49	83	33	231	117	1117
3 ノゲントンボ	0	0	0	0	0	3	0	0	0	0	0	18	19	24	18	17	99
3 モノサシトンボ	0	0	0	0	0	3	0	0	0	0	0	18	19	24	18	17	99
イトトンボ科	96	338	870	129	362	183	577	417	322	482	334	279	291	313	547	502	6022
4 アジアイトトンボ	17	244	595	35	74	40	57	24	25	107	49	8	23	56	154	107	1615
5 ルリイトトンボ	18	54	103	62	106	38	379	283	202	93	32	39	27	23	91	29	1579
6 クロイトトンボ	1	2	54	4	65	18	11	14	19	10	50	149	128	114	103	110	852
7 オオイトトンボ	4	1	67	6	37	6	8	3	4	10	97	3	1	10	62	128	447
8 セスジイトトンボ	49	31	22	9	25	1	3	18	0	0	4	75	111	102	48	52	550
9 オゼイトトンボ	0	0	0	1	0	1	2	3	0	0	1	0	0	0	1	0	9
10 エゾイトトンボ	5	3	10	5	48	55	83	48	24	29	24	4	1	6	69	71	485
11 キタイトトンボ	2	3	19	7	7	24	34	24	48	213	77	1	0	2	19	5	485
ヤンマ科	10	27	42	30	54	31	21	28	57	56	34	2	7	12	61	29	501
12 アオヤンマ	0	3	2	0	9	2	1	0	7	3	2	0	0	0	16	4	49
13 オオルリボシヤンマ	1	5	11	12	20	12	1	7	15	19	9	1	3	0	29	17	162
14 マダラヤンマ	2	3	15	8	5	5	4	7	16	11	13	0	1	3	10	5	108
15 ギンヤンマ	7	16	14	10	20	12	15	14	19	23	10	1	3	9	6	3	182
エゾトンボ科	0	7	1	1	11	3	4	4	9	0	6	0	1	1	2	3	53
16 オオトラフトンボ	0	6	1	1	11	3	4	4	9	0	6	0	1	1	2	3	52
17 カラカネトンボ	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
トンボ科	177	171	452	261	302	344	168	179	369	297	172	328	442	475	499	344	4980
18 ヨツボシトンボ	0	1	4	0	4	4	0	2	1	2	2	0	0	0	1	6	27
19 シオカラトンボ	44	31	100	32	53	49	5	46	58	36	17	47	53	149	18	23	761
20 ナツアカネ	1	2	9	6	8	11	1	14	7	15	19	1	3	4	5	3	109
21 アキアカネ	78	88	271	108	89	122	99	33	96	118	31	50	28	83	267	161	1722
22 マユタテアカネ	1	0	0	0	0	1	0	1	2	0	0	125	299	143	61	50	683
23 マイコアカネ	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	2	5
24 エゾアカネ	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
25 ヒメリスアカネ	0	0	0	0	12	17	0	0	3	0	1	0	2	0	3	0	38
26 ノシメトンボ	51	48	67	113	130	138	63	83	201	125	102	102	55	95	142	98	1613
27 キトンボ	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	3
28 ウスバキトンボ	2	1	1	1	5	1	0	0	0	0	0	3	1	1	1	1	18
合計 個体数	337	621	1454	521	882	755	954	753	893	950	655	681	850	880	1379	1059	13584
種数	18	21	20	20	22	25	19	20	21	18	21	18	21	19	25	23	28

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

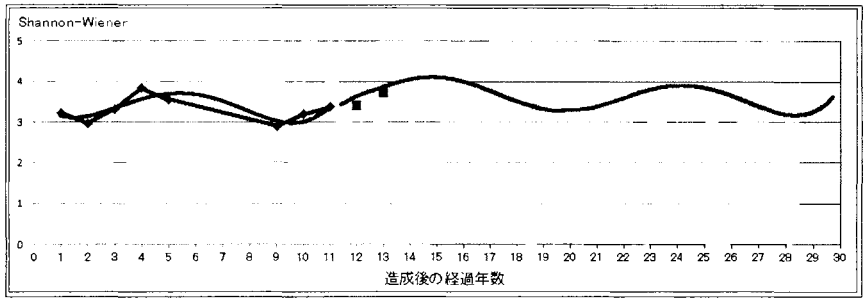
調査結果





8

調査結果



- この結果は日本学生科学賞の中央審査会で発表

この変動の要因となっている種を探そう！

9

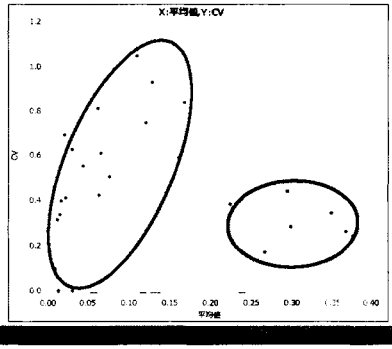
所有のデータ(トンボ相の定量調査の結果)

<解析方法(JMP)>

値の大きさと変動の様子を調べる

↓

① 種ごとに多様度指数のCV・平均値を計算



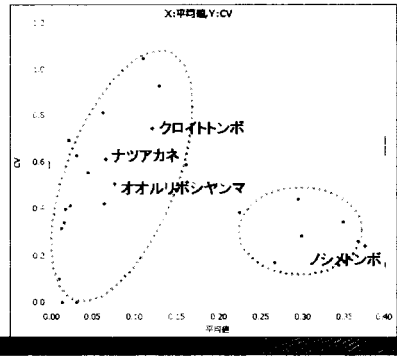
↓

2つにグループ分け

10

所有のデータ(トンボ相の定量調査の結果)

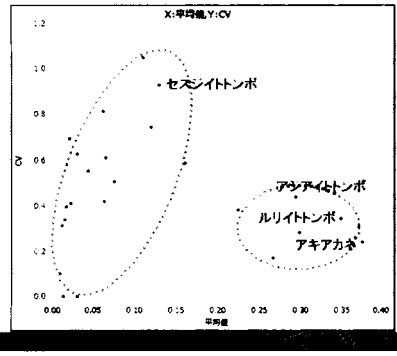
② 自由度とt値からp値を種ごとに計算し、
全体の多様度指数との正の相関を調べた (p<0.05)
⇒ ノシメトンボ、オオルリボシヤンマ、
クロイトトンボ、ナツアカネ に絞った



散布図では・・・
ノシメトンボだけが
右下のグループに！

所有のデータ(トンボ相の定量調査の結果)

③ 全体の多様度指数との負の相関を調べた
(p＞0.2)(r＜0)
⇒ アジアイトトンボ、ルリイトトンボ、
セスジイトトンボ、アキアカネ に絞った



散布図では・・・
セスジイトトンボ 以外
右下のグループに！

所有のデータ(トンボ相の定量調査の結果)

④ アジアイトトンボ、ルリイトトンボ、アキアカネ と、
全体の多様度指数との相関係数rを調べた

⇒ **アジアイトトンボ** が最小(-0.375)

⑤ アジアイトトンボ、ルリイトトンボ、アキアカネ と、
ノシメトンボとの相関係数rを調べた

⇒ **アジアイトトンボ** が最小(-0.663)

➡ **アジアイトトンボが最も関係がある**

1

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

所有のデータ(トンボ相の定量調査の結果)

<解析結果>

散布図の右下のグループ
→ 多様度指数の平均が高い・CVが小さい

⇒ 高い値で小さく変動している
●ノシメトンボ → 相関が最大
●アジアイトトンボ → 相関が最小

⇒ 全体の変動に大きく関係している
ノシメトンボ & アジアイトトンボ
の採集で全体の多様性変動がわかる?

14

所有のデータ(トンボ相の定量調査の結果)

〈解析結果から〉

ノシメトンボ や アジアイトトンボや 生物多様性についての研究は、この約40年間で、

何について、どれくらい 行われているのかを知るため、JSTの論文データをJMPで解析した

1

JSTの論文データ

〈解析方法(JMP)〉

① 和文表題の論文からトンボ関係の論文のみを抽出

② 抽出した論文数が全体の何%を占めるのか調べた

③ 奈良先端科学技術大学院大学(松本研究室)の「茶筌」で、抽出した論文タイトルを単語に分割

16

JSTの論文データ

〈解析方法(JMP)〉

④ トンボについての研究の経年変化や、
論文名を単語に分割した結果から、
トンボの研究で多いテーマを調べた

⑤「生物多様性」「ノシメトンボ,アカネ」
「アジアイトトンボ」が題に使用されている論文の
発行年を結合し、経年変化を見た

17

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

JSTの論文データ

〈解析結果〉

②「生物多様性」・・・257本 → 約0.008%

トンボ研究の論文数・・・172本 → 約0.005%

「ノシメトンボ,アカネ」・・・10本 → 約0.0003%

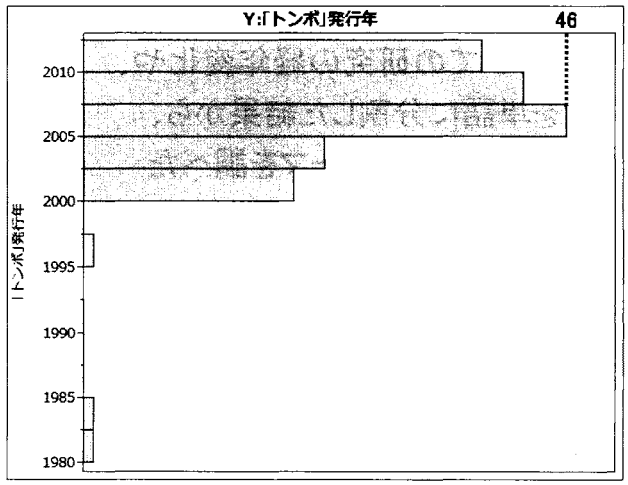
「アジアイトトンボ」・・・8本 → 約0.0002%

私達のような定量調査は他にない！

18

JSTの論文データ

④ トンボの研究は、2000年から急激に増えている



19

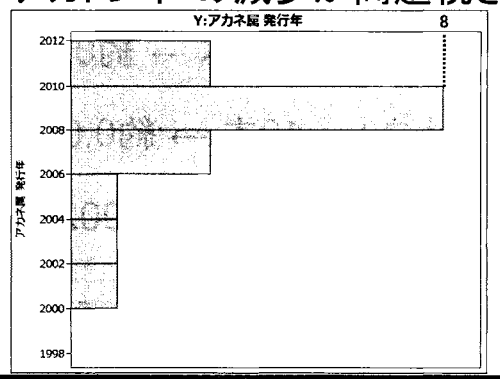
医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

JSTの論文データ

1997年　京都議定書 締結

→ 1990年後半～2000年前後

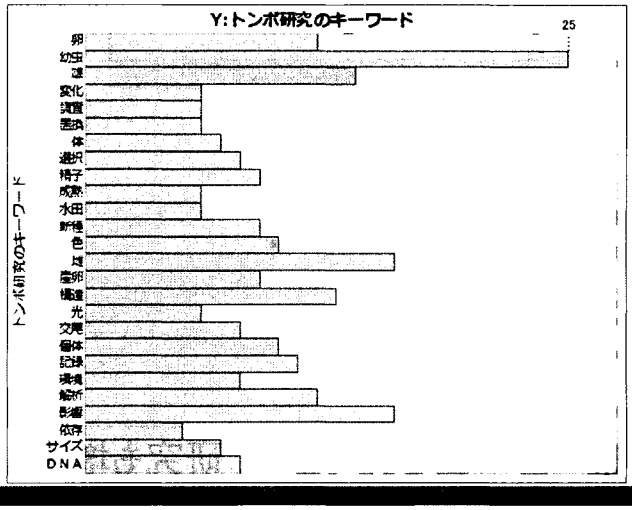
アカトンボ の減少が問題視され始めた



20

JSTの論文データ

④ トンボの研究では、幼虫(ヤゴ)に関することが多い

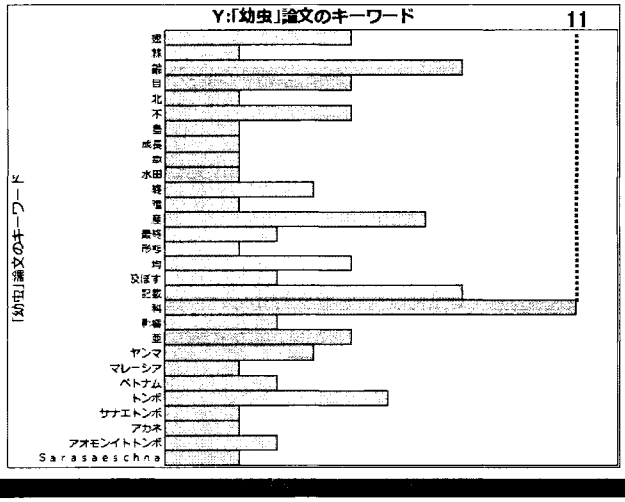


21

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

JSTの論文データ

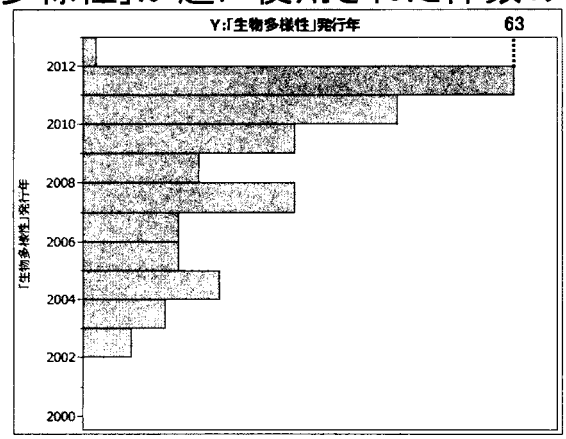
④ 幼虫の研究では、「科」を調べているものが多い



22

JSTの論文データ

⑤「生物多様性」が題に使用された件数の経年変化

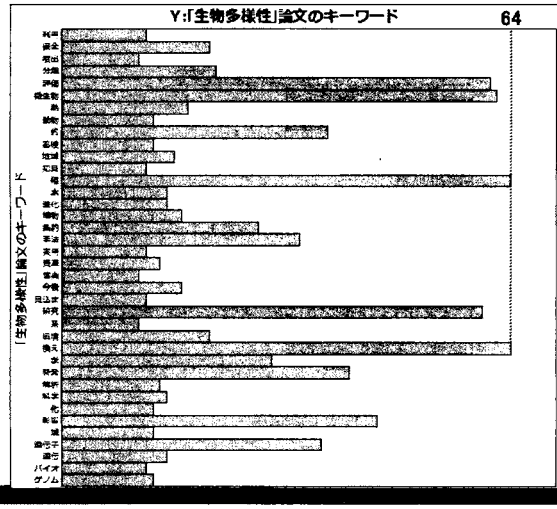


生物多様性が注目されてきて、研究も増えている 23

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

JSTの論文データ

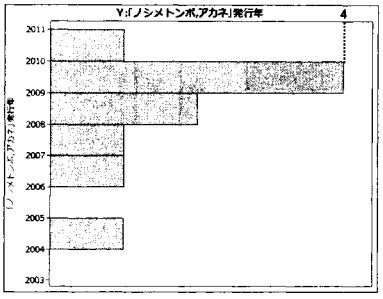
⑤「生物多様性」の研究では、遺伝子組換えに関することが多い

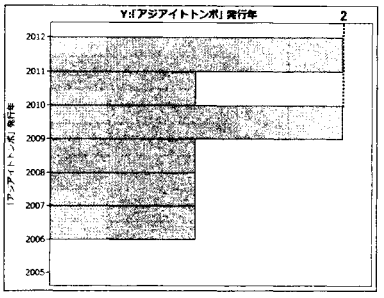


24

JSTの論文データ

⑤「ノシメトンボ,アカネ」「アジアイトトンボ」が
タイトルに使用された件数の経年変化





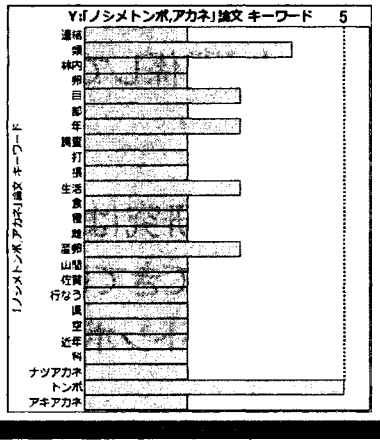
ノシメトンボ と アジアイトトンボの研究は、
まだまだ少ない

25

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

JSTの論文データ

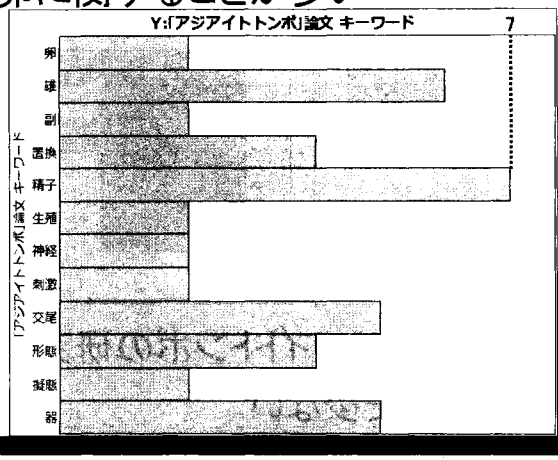
⑤「ノシメトンボ,アカネ」の研究では、
「トンボ目トンボ科アカネ類」の分類でのものが多い



26

JSTの論文データ

⑤「アジアイトトンボ」の研究では、
交尾産卵に関することが多い



まとめ

- 5年間のトンボ相の定量調査
  → ノシメトンボ と アジアイトトンボ が
  多様性変動に大きく関係している

- JSTタイプBのデータ
  → 生物多様性やトンボの研究は増えてきている
  → 最近になって注目されてきている
  → ノシメトンボやアジアイトトンボの研究は少ない

まとめ

- JSTタイプBのデータ
  - → トンボの研究・・・科ごとにヤゴ
  - → 生物多様性・・・遺伝子組換え
  - → ノシメトンボ・・・トンボ目トンボ科アカネ類
  - → アジアイトトンボ・・・交尾産卵

考察

- 生物多様性は重要であり、ノシメトンボとアジアイトトンボは多様性と大きく関わっている
- 今後はノシメトンボやアジアイトトンボの種としての特性や、生物多様性の研究を増やしていくべきである

参考

- 田久浩志ほか,2006,JMPによる統計解析入門第2版
- 「茶筌」:奈良先端科学技術大学院大学
  情報科学研究科 自然言語処理学講座（松本研究室）

謝辞

統計解析やJMPの使用方法などを教えていただいた
北海道大学 地球環境科学研究院 の片平浩孝先生、
的確なアドバイスを下さった顧問の綿路先生、
トンボの調査に協力していただいた方々に、
改めてお礼申し上げます

31

政府におけるオープンデータの取組について

鈴木 一広
内閣官房 / 情報通信技術(IT)総合戦略室 / 参事官

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

我が国のオープンデータの取組は、平成24年7月にIT戦略本部で「電子行政オープンデータ戦略」が決定されてから、具体的に進められてきました。

このオープンデータの取組について、平成26年6月に改定された「世界最先端IT国家創造宣言」等の基本方針の概要を説明するとともに、公開データの利用ルール(政府標準利用規約)の整備と府省横断的なデータ検索システム(データカタログサイト)に関する最近の具体的取組を説明します。また、今後の展望と課題についても説明します。

JST情報資産のオープン化について

佐藤 正樹

独立行政法人科学技術振興機構(JST)／
情報企画部海外・連携推進グループ／調査役

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

JSTの情報事業は昭和32年(1957年)にJICSTとして発足して以来、50数年にわたって国内外の科学技術文献(論文、学術雑誌記事、国や地方公共団体の研究報告書、企業技報など)を収集、整理・体系化してきました。
具体的には、国内資料12,000誌、海外資料4,700誌をもとに現在約3,000万論文の書誌情報を蓄積している他、300万件以上の化学物質の情報や100万語以上の科学技術・特許用語の体系等を整備しています。
作成したJST科学技術データは、これまで主に文献の書誌・抄録データベースから供され、先行技術調査をはじめとする検索のために広く使用されてきましたが、昨今は意思決定を支援するための分析や言語処理的な研究などに利用したいというニーズに応えるため、バルクデータ(CSV,JSON)やWeb経由(RDF)による提供体制も整備しています。
昨年度は、今般のデータオープン化の流れを受け、分析に資するJSTデータの提供を加速するための足がかりとして、SAS社と共催で統計・分析コンテスト「データサイエンス・アドベンチャー杯(http://www.sascom.jp/AAC/)」を開催しました。このコンテストを行ったことでJST科学技術データに対するニーズが把握できたことから、まずは研究用途でのデータオープン化を推進することになりました。
本発表では、JSTの保有する科学技術データの説明を行うとともに、データ活用の事例紹介を交えたJST情報事業のデータオープン化の取組みについてお話しします。

2

Let's 匿名データ分析: パック旅行費支出と世帯情報の関連の検討

○魚住 龍史
京都大学大学院医学研究科 医学統計生物情報学

Let's analyze anonymous data: investigation for the relationship between travel fees and household characteristics

Ryuji Uozumi
Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine

要旨

旅行・観光市場は世界的にみて成長性の高い産業であり，波及効果の裾野の広い産業である．特に，各旅行会社は高齢者向けのサービスを多く提供している．魚住 (2013) の報告によると，全国消費実態調査の個票データから作成された，擬似ミクロデータを用いたデータ分析の結果，パック旅行費支出額は年齢とともに増加し，特に 2 人世帯において支出額が高いことが示唆された．本稿では，擬似ミクロデータよりも多くの情報が含まれた匿名データを用いて，パック旅行費支出と世帯情報の関連を検討したデータ分析結果を報告する．

キーワード：匿名データ，データ分析，パック旅行費，国内旅行費，外国旅行費，10 大費目，教養娯楽，世帯情報，集計用乗率

1 はじめに

旅行・観光市場は世界的にみて成長性の高い産業であり，波及効果の裾野の広い産業である．観光庁において，旅行・観光産業の経済効果に関する調査研究が定期的に実施されている．この調査は，日本の旅行・観光における消費実態を明らかにし，旅行・観光施策の基礎資料のために活用することを目的として実施されている．2012 年の調査によると，国内旅行消費は 22.5 兆円であり，世界的に見ても上位の国として位置づけられる．さらに，国内旅行消費が日本経済へ及ぼす影響としても，生産波及効果は 46.7 兆円といわれており，雇用創出効果は 399 万人であると報告されている [1].

SAS ユーザー総会では，2013 年より Let's データ分析コンテストが行われている．このコンテストは規定課題と自由課題に取り組むことが課せられている．魚住 (2013) による報告では，平成 16 年全国消費実態調査の個票データに基づいて作成された擬似ミクロデータを用いて，パック旅行費支出と世帯情報の関連の検討が行われた [2–4]．結果，パック旅行費は年齢とともに増加し，2 人世帯ほどパック旅行費への支出額が高いことが示された．しかし，擬似ミクロデータでは，パック旅行費が国内旅行と外国旅行のどちらのデータであるか判別できず，さらなる検討はできなかった．

本稿では，平成16年全国消費実態調査の匿名データを用いて，パック旅行費支出と世帯情報の関連を検討することを目的とする．また，その他の家計簿における収入・支出データを用いて，パック旅行費との関連を検討する．なお，本稿で用いる匿名データは，独立行政法人統計センターによって作成され，一般の利用に供することを目的として調査票情報を特定の個人または法人その他の団体の識別 (他の情報との照合による識別を含む) ができないように加工されている [5]．データとしては，1780変数格納されたCSV形式のファイルであり，単身世帯は3,936行，2人以上の世帯は43,861行のデータとなる．なお，8人以上の世帯はデータから削除されている．さらに，本データには集計用乗率が含まれており，家計収支編に使われた乗率をリサンプリングの抽出率で調整したものである．集計用乗率を重みとすると，単身世帯で873,232件，2人以上の世帯で335,797件のデータとなる．なお，リサンプリングは単身世帯と2人以上の世帯それぞれにおいて，地域ごとに調査報告書に掲載されている結果表のうち，家計収支編等に使われた基本的な集計時に用いた集計用乗率の大きさに基づいて層化した上で，各層において抽出率が約8割となるように集計用乗率による確率比例抽出を行っている．本データをSASにインポートした上で，検討を行う．

2 パック旅行費支出と世帯情報の関連の検討

本稿では，パック旅行費に焦点を当てる．匿名データにおいて，パック旅行費は消費支出における大分類「教養娯楽」，中分類「教養娯楽サービス」に含まれる変数である．

本稿で用いる要約指標としては，集計用乗率で重み付けした平均値を用いる．まず，匿名データにおいて，パック旅行費に影響を及ぼす可能性のある世帯情報として，年齢カテゴリ，世帯人員，企業区分，企業規模，産業カテゴリ，職業カテゴリが挙げられる．これらを説明変数として，GLMSELECTプロシジャにより変数選択を行った．なお，パック旅行費は，支出していない世帯については0となり，結果，図1-1のような右にスソを引いた分布となる．そこで，図1-2のような対数変換後のパック旅行費を目的変数として用いた．変数選択はSTEPWISE法を採用し，Schwarz Bayesian information Criteria に基づき実施した．結果，年齢カテゴリ，世帯人員，企業規模が選択された．以下，これらの要因がパック旅行費にどのように影響を及ぼすか検討する．なお，年齢カテゴリは25歳未満，75歳以上のカテゴリ化，25歳以上75歳未満については5歳刻みでカテゴリ化した，計12カテゴリの変数である．

なお，本検討では，勤労世帯のみを抽出し，単身世帯及び2人以上7人以下の世帯の匿名データ27,443オブザベーションを対象とした．

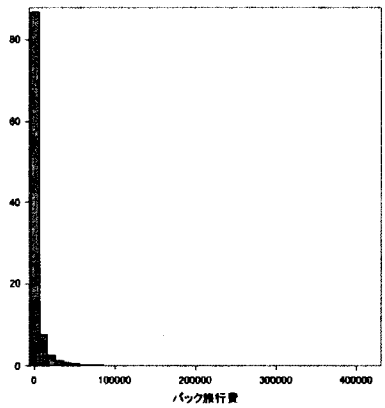


図1-1：パック旅行費の分布 (円)

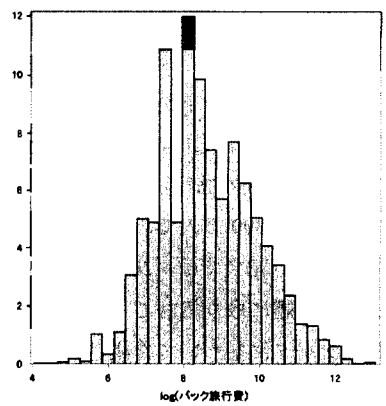


図 1-2：対数変換後のパック旅行費の分布

2.1 年齢による影響

各年齢における 10 大費目の支出割合を図 2-1，各年齢におけるパック旅行費の支出額 (円) を図 2-2，各年齢における教養娯楽の支出割合を図 2-3 に示す. 図 2-2 における支出額は集計用乗率による重み付き平均を示している．なお，図 2-1 と図 2-3 は GCHART プロシジャ，図 2-2 は SGPLOT プロシジャを用いて作成した.

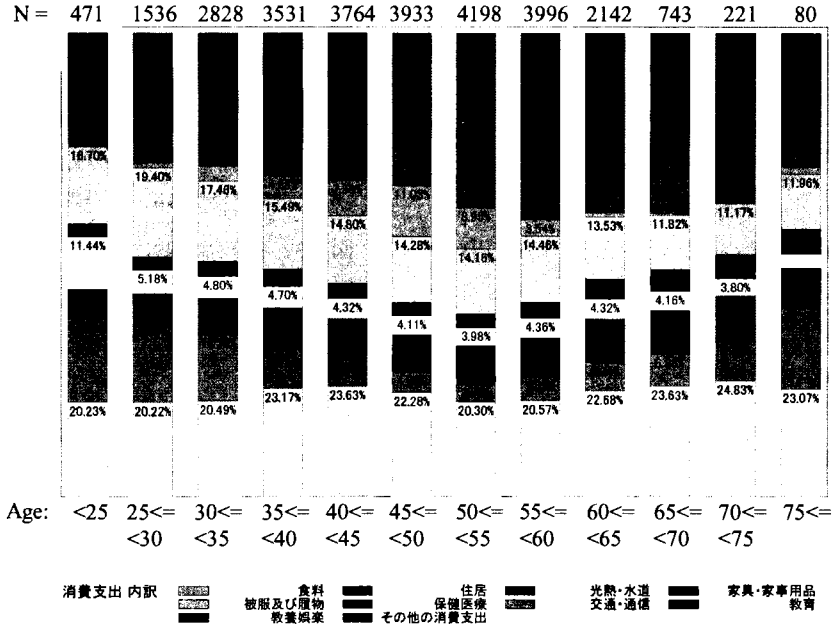


図2-1：各年齢における10大費目の支出割合

図 2-1 は，下から食料，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順に積まれたグラフである.

図 2-1 より，例えば，消費支出に占める住居の割合に注目すると，75 歳以上のカテゴリを除いて，年齢とともに支出は減少している．また，消費支出に占める交通・通信の割合においても，年齢とともに支出は減少している．しかし，今回焦点を当てるパック旅行費が含まれている教養娯楽に着目すると，年齢が変化しても支出は横ばいとなっている．そこで，各年齢におけるパック旅行費の支出額のグラフとして，図 2-2 を見て検討する.

図 2-2 より，パック旅行費は年齢とともに増加していることが分かる．特に，年齢が 55 歳以上のグループにおいて，パック旅行費は顕著に高くなっている．さらに，図 2-3 において，各年齢における教養娯楽に占めるパック旅行費の支出割合を見ても，年齢とともにパック旅行費支出割合は高くなっていることが分かる.

ここで，図 2-1 において，消費支出に占める教育の割合が低い年齢では，図 2-2 及び図 2-3 におけるパック旅行費の支出額及び支出割合が高くなっていることが分かる．本データにおいて，10 大費目の 1 つである教育費には教科書・学習参考教材，補習教育，授業料等が含まれている．すなわち，子供の授業料，教科書・学習参考教材，補習教育による教育費への支出が多い年齢の世帯においては，パック旅行費に支出できる割合が低くなってしまうと考えられる.

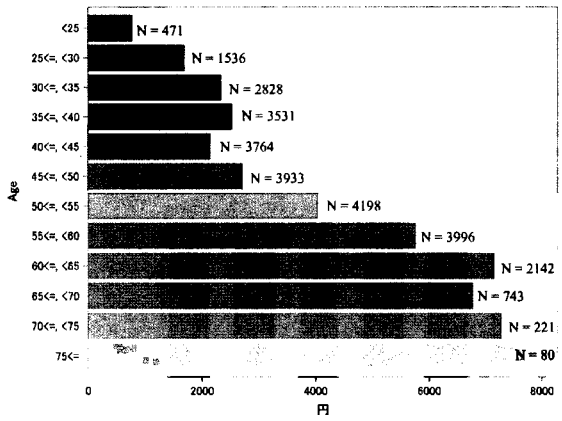


図2-2：各年齢におけるパック旅行費支出額 (円)

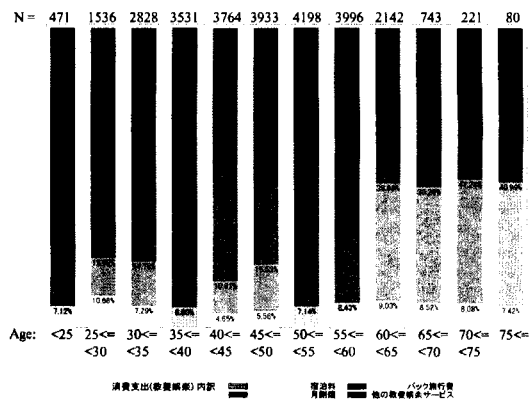


図 2-3：各年齢における教養娯楽の支出割合

さらに，各年齢におけるパック旅行費の支出額を，国内旅行費，外国旅行費に分けて図 2-4 に示した．なお，図 2-4 は SGPANEL プロシジャを用いて作成した．

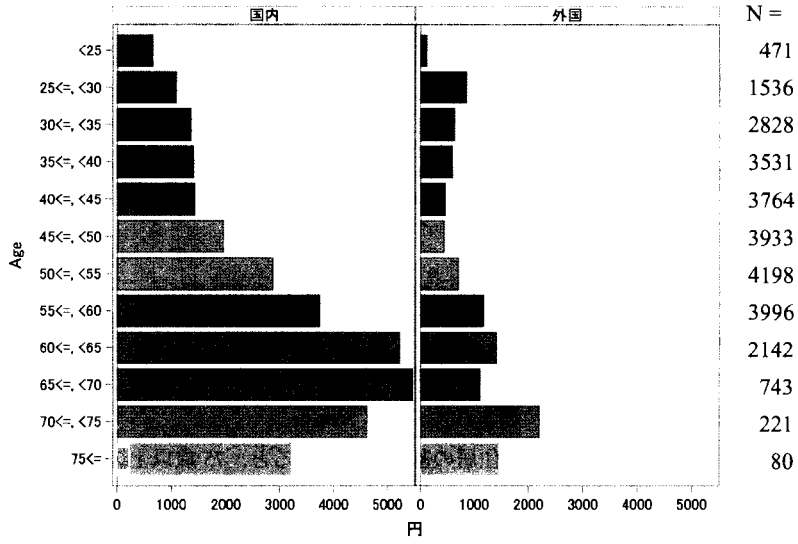


図2-4：各年齢における (国内 / 外国) パック旅行費支出額 (円)

図 2-4 より，パック旅行費の支出が多い年齢が 55 歳以上のグループは，主に国内パック旅行として利用していることが分かる．一方，外国パック旅行への支出は，国内パック旅行費に比べて，そこまで多くないことが分かる．

2.2 世帯人員による影響

各世帯人員における 10 大費目の支出割合を図 3-1，各世帯人員におけるパック旅行費の支出額を図 3-2，各世帯人員における教養娯楽の支出割合を図 3-3 に示す．

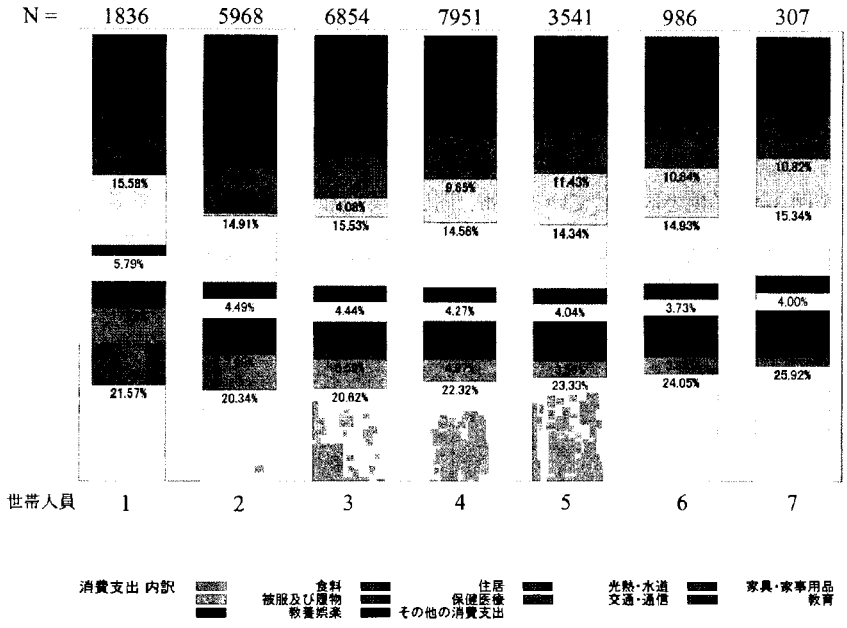


図 3-1：各世帯人員における 10 大費目の支出割合

図 3-1 より，パック旅行費が含まれている教養娯楽について，単身世帯では支出割合が若干高いが，2 人以上の世帯においては世帯人員数依らず支出割合は横ばいとなっている．しかし，教育に注目すると，特に 4 人以上の世帯においては，消費支出に占める教育の割合が高くなっている．この背景として，世帯人数が多いことによって，子供がいる世帯であることが考えられる．その結果，子供の授業料，教科書・学習参考教材，補習教育による教育費の占める割合が高くなったと考えられる．

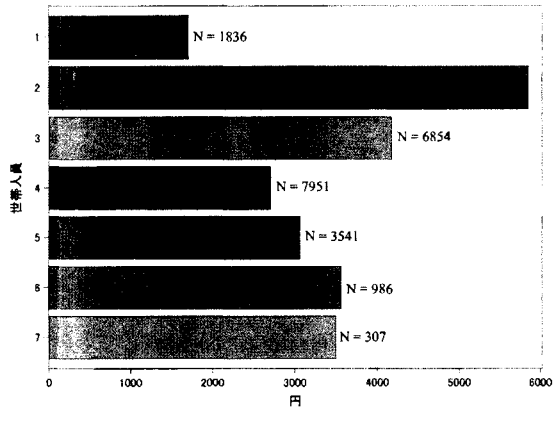


図3-2：各世帯人員におけるパック旅行費支出額 (円)

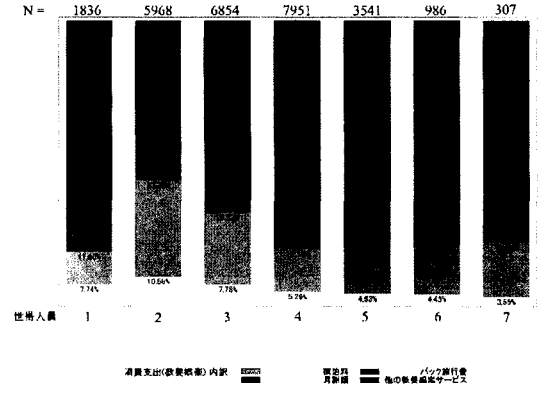


図 3-3：各世帯人員における教養娯楽の支出割合

次に，図 3-2 及び図 3-3 より，特に 2 人世帯において，パック旅行費の支出額及び教養娯楽に占めるパック旅行費の支出割合は高くなっていることが分かる．ここで，図 3-1 において，単身世帯及び 2 人世帯で 10 大費目に占める教育の割合が低いことが分かる．消費支出に占める教育の割合の低い 2 人世帯においては，その分パック旅行費に支出できる割合が高くなったと考えられる．

上述のように，子供の教育への支出がパック旅行費支出に影響しているか検討するために，さらなるデータ分析として，子供と同居している世帯 (生計が同一であることを意味する)，同居していない世帯別のパック旅行費支出額を図 3-4，子供と同居している世帯，同居していない世帯別の教養娯楽の支出割合を図 3-5 に示す．

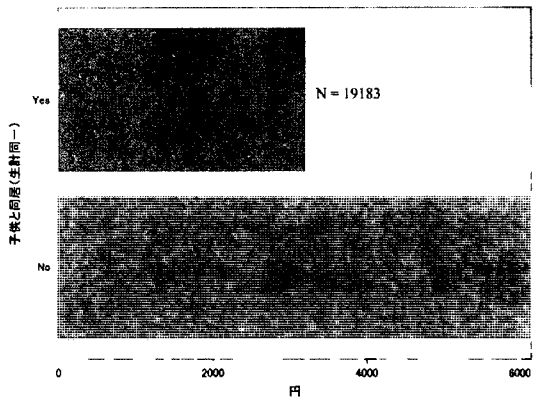


図3-4：各世帯特徴におけるパック旅行費支出額 (円)

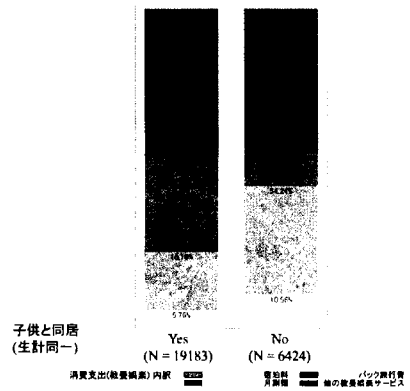


図 3-5：各世帯特徴における教養娯楽の支出割合

図 3-4 より，パック旅行費支出額は子供と同居している世帯の方が少ないことが分かる．また，図 3-5 より，教養娯楽に占めるパック旅行費の割合は，子供と同居している世帯の方が低いことが分かる．

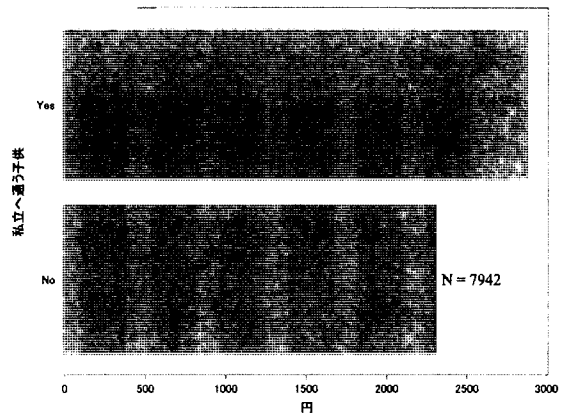


図3-6：各世帯特徴におけるパック旅行費支出額 (円)

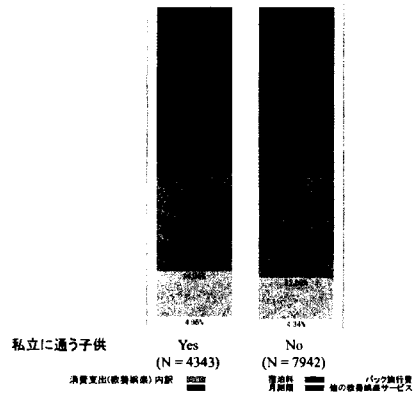


図 3-7：各世帯特徴における教養娯楽の支出割合

図 3-4 及び図 3-5 より，子供との同居の有無がパック旅行費支出に影響を与えることが分かった．さらに，子供の通学先の学校形態を検討する．本匿名データでは，子供が私立あるいは国公立のどちらの形態の学校

に通学しているかという情報が含まれている．そこで，学費がより多くかかることが予想される，私立へ通う子供の有無別のパック旅行費支出額を図 3-6，私立へ通う子供の有無別の教養娯楽の支出割合を図 3-7 に示す．なお，私立へ通う子供 "No" は，国公立へ通う子供の "Yes" を意味している．

図 3-6 より，私立へ通う子供がいる世帯の方が，パック旅行費支出額が高いことが分かる．これは，私立へ通う子供がいる世帯の方が，国公立へ通う子供がいる世帯に比べて，経済的に余裕のあることが背景として考えられる．次に，図 3-7 より，私立へ通う子供がいる世帯といない世帯で，教養娯楽の支出割合は同程度であることが分かる．以上より，私立へ通う子供がいる世帯の方が，国公立へ通う子供がいる世帯に比べて経済的に余裕があり，パック旅行費に限らず，消費支出自体が多いことが分かる．

2.3 2 人世帯に対する検討

2.2 において，2 人世帯のパック旅行費支出額が高いことが分かった．そこで，さらなる検討として，2 人世帯を対象として，**2.1** で検討した，年齢による影響についてデータ分析する．

まず，2 人世帯を対象とした，各年齢における 10 大費目の支出割合を図 4-1，各年齢におけるパック旅行費の支出額を図 4-2，各年齢における教養娯楽の支出割合を図 4-3 に示す．

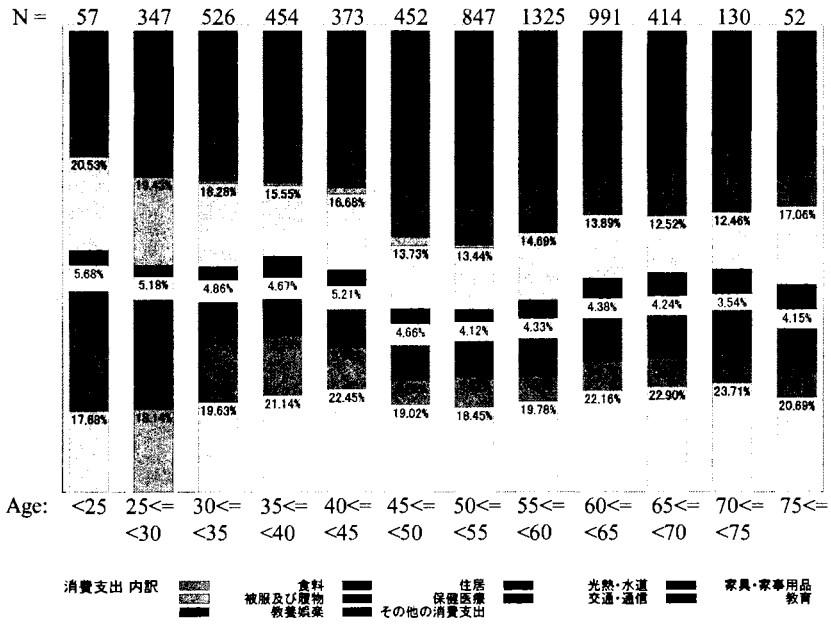


図4-1：各年齢における10大費目の支出割合

図 4-1 は，図 2-1 と比べて，10 大費目に占める教育の割合が極めて低くなっていることが分かる．また，その他の消費支出の割合が高くなっている．ここで，本データのその他の消費支出としては，諸雑費 (理美容サービス，理美容用品，身の回り用品，たばこ，小遣い (使途不明))，交際費 (食料，家具・家事用品，被服及び履物，教養娯楽，他の物品サービス，贈与金)，仕送り金等が含まれる．これは，パック旅行費と同じように，比較的経済的に余裕のある世帯において，支出が多くなる項目であると考えられる．よって， 2 人

世帯を対象にした図 4-1 において，消費支出に占める教育の割合の低い世帯では，その他消費支出に支出できる割合が高くなったと考えられる．

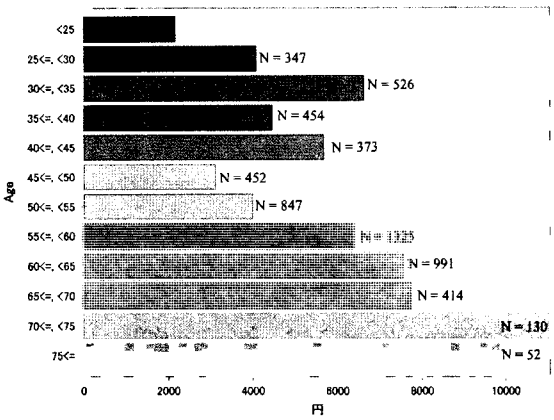


図4-2：各年齢におけるパック旅行費支出額 (円)

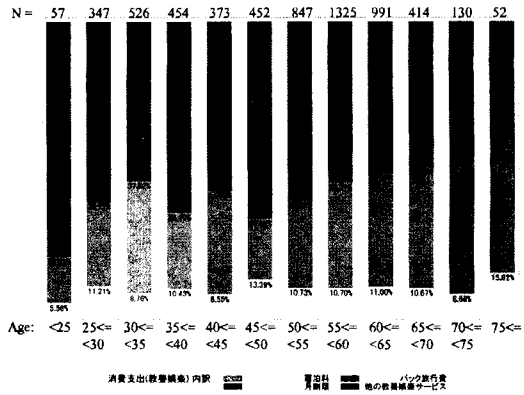


図 4-3：各年齢における教養娯楽の支出割合

図 4-2 及び図 4-3 では，単身世帯から 7 人世帯までを対象とした図 2-2 と異なり，年齢とともにパック旅行費支出額は単調に増加しておらず，30 歳代の世帯においても支出額が比較的高い，2 峰性の分布となっていることが分かる．さらに，60 歳以上の世帯においては，図 2-2 と比べて，パック旅行費支出額が高くなっている．図 2-2 及び図 2-3 においては，年齢が高くなるほどパック旅行費支出額は高くなることが示唆された．しかし，2 人世帯を対象とした検討を行った結果，年齢が高いだけがパック旅行費支出に寄与していないことが考察された．

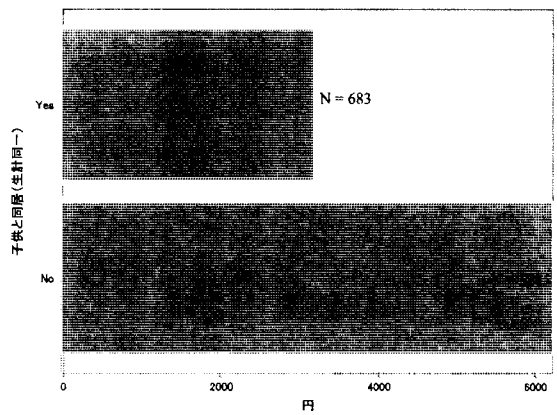


図4-4：各世帯特徴におけるパック旅行費支出額 (円)

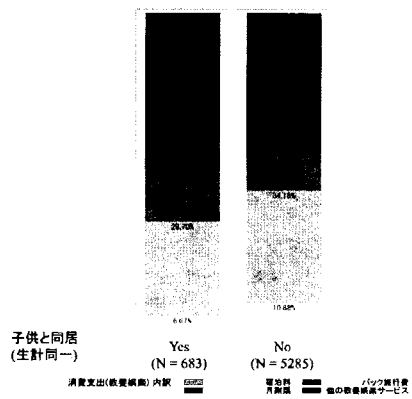


図 4-5：各世帯特徴における教養娯楽の支出割合

また，図 3-4 及び図 4-4 より，子供と同居している世帯のうち，2 人世帯である割合は 683 / 19183 = 3.6% であることが分かる．

以上より，2人世帯では子供と同居している世帯が少ないことによって，教育費に支出する割合が低くなり，その結果，パック旅行費支出額が全体的に高いことが分かった．

さらに，2人世帯を対象として，各年齢におけるパック旅行費の支出額を国内旅行費，外国旅行費に分けて図4-6に示した．

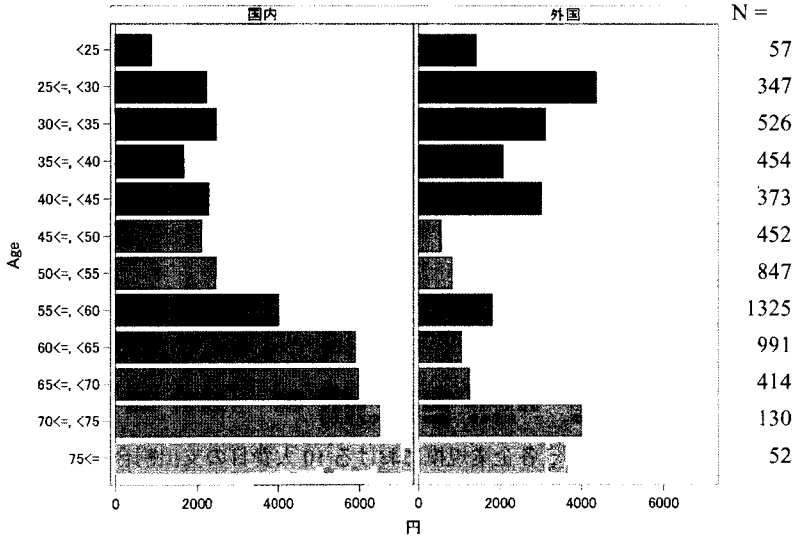


図4-6：各年齢における (国内 / 外国) パック旅行費支出額 (円)

図4-6より，単身世帯から7人世帯を対象として検討した図2-4と同様に，パック旅行費の支出が多い年齢が55歳以上のグループは，主に国内パック旅行として利用していることが分かる．一方，外国パック旅行費については，どの年齢増においてもそこまで大きな違いが見られなかった図2-4とは異なる分布が得られている．70歳以上及び25歳から45歳までの年齢において，外国パック旅行費支出額が高い傾向にあることが分かる．特に，25歳から35歳までの年齢において，外国パック旅行費支出額が高い要因として，婚姻件数の割合が高い年齢であることが挙げられる [6]．その結果，大多数の新婚夫婦が新婚旅行として，外国パック旅行を使うことが多い年代であると考えられる．

2.3 企業規模による影響

各企業規模における10大費目の支出割合を図5-1，各企業規模におけるパック旅行費の支出額を図5-2，各企業規模における教養娯楽の支出割合を図5-3に示す．

図5-1より，企業規模によって，パック旅行費が含まれている教養娯楽の支出割合が大きく異なることはないことが分かる．また，他の大費目についても，企業規模によって異なる項目はなかった．さらに，図5-3においても，企業規模によって，教養娯楽に占めるパック旅行費支出割合が異なることはなかった．

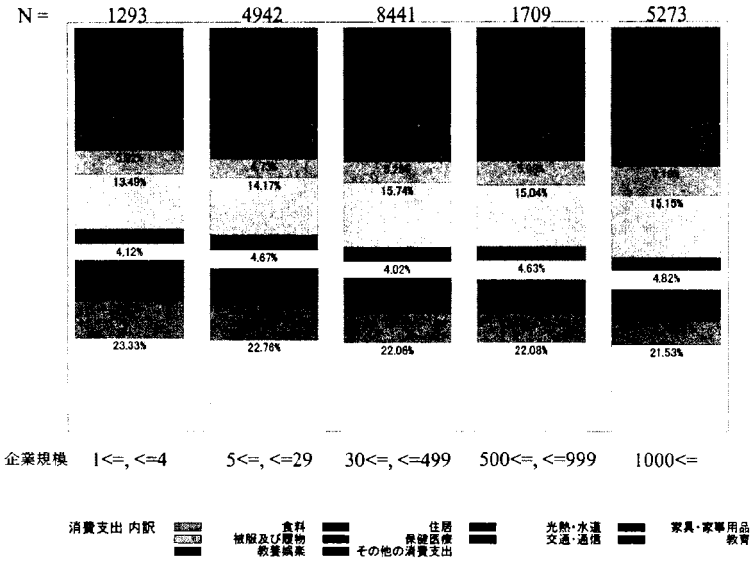


図 5-1：各企業規模における 10 大費目の支出割合

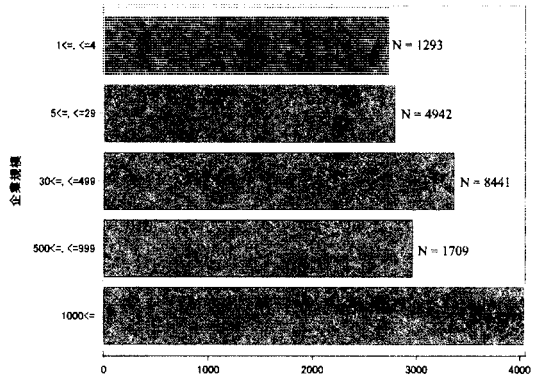


図5-2：各企業規模におけるパック旅行費支出額 (円)

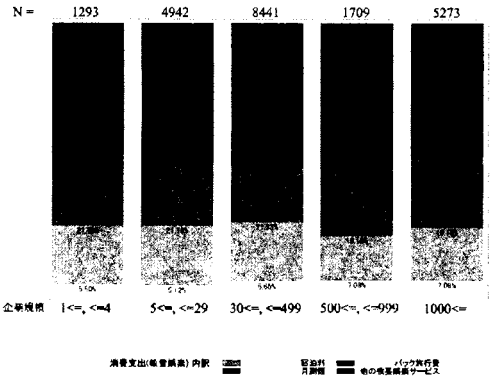


図 5-3：各職業における教養娯楽の支出割合

図 5-2 より，企業規模が大きくなるほどパック旅行費支出額は若干高くなっていることが分かる．これは，企業規模が小さい企業においては，パック旅行中の仕事の代理ができない等のことが起因していると考えられる．

3 まとめ

本稿では，パック旅行費に影響を与える世帯の特徴を表す要因として，年齢カテゴリ，世帯人員，企業規模がどのように影響を与えるか検討を行った．結果，パック旅行費支出は年齢とともに増加し，2 人世帯ほどパック旅行費への支出額が高いことが分かった．2 人世帯においてパック旅行費の支出額が高い要因として，子供の教育費に占める支出割合が低い世帯であることが考えられ，結果，パック旅行費へ割り当てることのできる支出割合が高くなったと考えられる．同様の議論より，子供と別居している世帯ほどパック旅行費の支出額が高い傾向にあることが分かった．また，子供がいる世帯においては，私立に通う子供を持つ世帯の方が，国公立へ通う子供を持つ世帯に比べて，パック旅行費の支出額が高い傾向にあった．さらに，パック旅行費の区分として，国内パック旅行と外国パック旅行に分けて検討を行った結果，全体的に年齢が高い世帯において支出額が高い傾向にあった．しかし，2 人世帯を対象とした検討の結果，外国パック旅行については，25 歳から 45 歳の年齢においても支出額が高いことが分かり，その要因として，婚姻に伴う外国への新婚旅行が考えられる．なお，2 人世帯を対象としたパック旅行費支出額の分布は 2 峰性を示し，外国パック旅行費支出の多い 25 歳から 45 歳の年齢層による影響が強いことが示唆された．ここで，年齢が高い世帯ほどパック旅行費の支出が多い想定の下，高齢者でも安心して旅行ができるバリアフリー特集，高齢者を対象とした人に優しい宿，60 歳からの旅行応援特集など，各旅行会社及び各団体は多くの高齢者向けのサービスを提供している状況である [7-10]．今後は，世帯において子供と同居しているかどうか調査した上で，子供と同居していない世帯や，子供のいない 2 人世帯をターゲットとしたプロモーションも，国内・外国旅行をより活性化させることにつながると期待される．

参考文献

[1] 観光庁. 旅行・観光産業の経済効果に関する調査研究 (2012 年版). http://www.mlit.go.jp/common/001040524.pdf. 2014/06/21 アクセス.

[2] 魚住龍史. ミクロデータ分析コンテスト: 規定課題. SAS ユーザー総会 論文集 2013, 507–510.

[3] 魚住龍史. 擬似ミクロデータによる国内旅行費支出と世帯情報の関連の検討. SAS ユーザー総会 論文集 2013, 511–514.

[4] 魚住龍史. SAS による擬似ミクロデータを用いたパック旅行費支出と世帯情報の関連の検討. 公的ミクロデータの利用に関する研究集会 2013.

[5] 独立行政法人統計センター. 匿名データ利用の手引 (学術研究・高等教育目的関係) 2013 年最終改正. http://www.nstac.go.jp/services/2ji/tokumei_tebiki.pdf. 2014/06/21 アクセス.

[6] 厚生労働省. 平成 22 年人口動態統計月報年計（概数）の概況. http://www.mhlw.go.jp/toukei/saikin/hw/jinkou/geppo/nengai10/kekka04.html. 2014/06/21 アクセス.

[7] H.I.S. 60 歳からの旅行を応援します. http://www.his-j.com/tyo/senior/. 2014/06/21 アクセス.

[8] JTB. バリアフリー特集. http://dom.jtb.co.jp/yado/theme/barrierfree/. 2014/06/21 アクセス.

[9] 全旅連 (全国旅館ホテル生活衛生同業組合連合会) シルバースター部会. 人に優しい宿. http://yadonet2.jp/. 2014/06/21 アクセス.

[10] 楽天トラベル. 人に優しい宿. http://travel.rakuten.co.jp/special/yasashii/. 2014/06/21 アクセス.

全国消費実態調査の匿名データを用いた、2人以上世帯の保険需要の分析

宇野　慧
アステラス製薬株式会社　開発本部　データサイエンス部

Analysis of households' (two people or more) insurance demand, by using National survey of family income and expenditure data.

Satoshi Uno
Data Science, Global Development, Astellas Pharma Inc.

要旨

保険商品は貯蓄型と非貯蓄型に大別されるが、世帯の加入行動の違いに関して分析を行った事例は少ない。本稿では平成16年度の総務省全国消費実態調査の匿名データを用い、貯蓄型保険と非貯蓄型保険両方の需要に対して、世帯属性の中でも特に就業状況が与える影響に着目した。推定モデルには、連立方程式の誤差の分散構造のみに関係性を盛り込んだSUR(Seemingly Unrelated Regression) Tobitモデルを用いた。その結果、就業状況が保険需要に与える影響は貯蓄型と非貯蓄型で大きく異なることが確認できた。特に、自営業世帯は非貯蓄型保険に対して非常に高い保険需要を示す一方で、貯蓄型保険に対する保険需要は非常に低い水準となることが分かった。

キーワード：全国消費実態調査匿名データ、保険需要、SUR Tobitモデル、NLMIXEDプロシジャ

1. はじめに

世帯の保険需要は、資産選択の問題として多くの実証分析の蓄積がある。濵本(2001)や岩本(2003)では「生命保険に関する全国消費実態調査データ」を用いて保険需要を分析している。このデータは公益財団法人生命保険文化センターが実施しているもので、保険支出に関する詳細な情報が調査されている。保険加入理由などの意識調査に関する質問項目がある一方で、保険以外の資産保有額については情報が少ない。

濵本(2001)では世帯主の保険金額、および家族全員の保険金額を被説明変数に、世帯主年齢、世帯主職業、子供の有無などを説明変数に設定した重回帰分析を行っている。その結果として、同居する子供の数や、自営業世帯であることが保険需要を高める要因であると指摘している。前者については、万一の時の生活保障金額がより大きく必要である事。後者については、世帯主が死亡した場合の影響が一般従業者世帯よりも大きい事が原因であると考察している。また岩本(2003)では、保険の加入動機についても着目した分析を行っている。具体的には、「満期保険受取額/老後に必要な資金総額」をライフサイクル目的の貯蓄動機の指標と定義している。貯蓄動機の強さが保険需要に与える影響をコントロールするために、死亡保険金額・満期保険

金額それぞれとの同時推定を行うType III Tobitモデルを構築し、貯蓄動機と保険需要を推定している。分析の結果、死亡保険金需要と満期保険金需要では世帯年収や資産保有額、住宅ローンの有無で影響の大きさが異なる傾向が見られることを示している。

その他のデータを用いた研究事例として、日経NEEDS　RADERデータを用いた浅野(1998)がある。浅野(1998)では個人年金と生命保険需要に対して、自らの老後に備えるライフサイクル動機、遺された子孫の生活を重視する遺贈動機が影響していると想定し、公的年金制度変更がこれらの動機に与えた影響について考察している。分析の結果、基礎年金支給開始年齢の引き上げに伴い、世帯主年齢が30代の世帯で生命保険を増額させる傾向が見られたことから、遺産動機による保険需要の可能性を示唆している。

以上の先行研究では、保険加入あるいは需要金額の分析として、家計の保険加入動機(ライフサイクル動機か、遺産動機かなど)や、あるいは機能面(死亡、疾病、老齢、それぞれ直面するリスクに対する個別需要)に着目したものが多い。しかしながら、商品特性としての貯蓄型/非貯蓄型の選択、という側面に関して明示的に分析を行っている研究は確認できなかった。

良く知られた事実として、保険商品は貯蓄型と非貯蓄型の2種類に大別することができる。一般的に貯蓄型保険では、保険料を継続的に支払い、契約期間を満了すると満期保険金として一括金額が受領できるタイプの保険が多い。また、契約期間中に解約した場合に解約返戻金が設定されていることから、これを満了前の貯蓄と考えることもできる。しかしながら、この解約金を通常の金融資産として扱うことは一般的ではない。満期時点での給付金額が決まっている確定給付型の保険が多く、原則的に満期までの長期的な資産運用の性質を持つ金融商品として、非貯蓄型と比べて月当たりの支払額が高額に設定されている点が特徴である。したがって、資産運用に関してリスク回避的な世帯の加入率が高く、支払額も高いと考えられる。

一方で非貯蓄型保険は毎月の掛け捨て型であるため、貯蓄型のような金融資産的な性質は持たない。月当たりの支払額は貯蓄型に比べて安価に設定されていることが多い。そのため、リスクを選好する世帯では、非貯蓄型保険で支出を安価に抑え、その差額分を別の金融商品として資産運用に回す可能性が考えられる。以上のように、貯蓄型と非貯蓄型では保険商品の特性も異なり、世帯の各保険に対する需要も大きく異なると考えられる。そのため本稿では、各保険の加入行動に影響を与えている世帯属性に着目して分析を行った。本稿で用いたデータは、平成16年度の全国消費実態調査の匿名データのうち、世帯員が2人以上の43861世帯である。匿名データでは調査世帯が特定されないために、貯蓄残高など一部の変数についてのトップコーディングや、リサンプリングなどの一部加工を行っている。そのため、実際の調査の集計結果と完全には一致しない点に注意が必要である。本稿では、提供されたデータをそのまま用いて分析を行った。

以下で本稿の構成を概説する。2節では、分析に用いたデータの変数名や作成方法などの詳細について説明する。また、非貯蓄型保険、貯蓄型保険それぞれの保険料支払額について、基本統計量と度数分布図により分布を確認する。3節では、本稿で行う分析に関する理論的な背景を説明すると共に、推定に用いたプログラムの概要を記載する。詳細なプログラム本文については、別途確認されたい。4節では、推定結果の概略を説明し、加えて簡単な解釈を行う。5節では、本稿のまとめと、今後の分析に対する展望を述べる。

2. データセットの特性について

2-1. 分析に用いた変数名、および作成方法

前節で述べた通り、本稿では平成16年度の全国消費実態調査の匿名データを用いて分析を行う。今回分析対象とした「世帯員が2人以上の世帯」については、平成16年度の9月から11月の3か月間の消費支出が

調査されている。そのため多くの世帯でボーナス月の影響が除かれており、年間全体を通じた支出を表しているわけではない点に注意が必要である。本調査の特徴として、各世帯人員の年代や就業状況等についても詳細に調査を行っている。また、貯蓄残高などストックのデータもあり、全てで1780系列のデータセットとなっている。なお、今回のデータでは貯蓄性保険のうち年金保険は他と性質が異なり、かつ加入している世帯が僅少であるため分析からは除外している。また、分析の対象としている貯蓄性保険と非貯蓄型保険については、自動車保険や住宅用火災保険は別項目として計上されている。そのため、支出額の詳細な内訳は不明であるが、生命保険(疾病特約を含む)および医療保険が支出の大半であると考えられる。本稿の分析で用いた変数の情報について、以下の表2-1にまとめた。

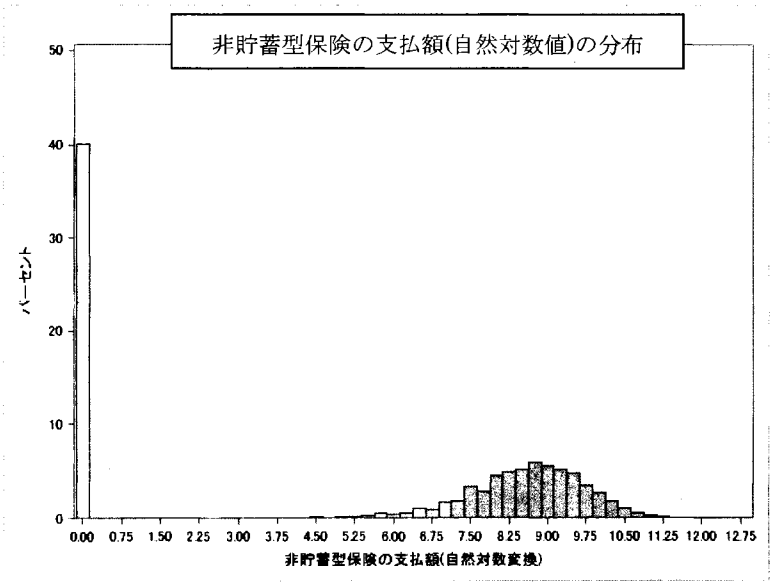
表2-1　分析に用いた変数一覧

変数名	区分	詳細説明
非貯蓄型保険	被説明変数	非貯蓄型保険料の支払い額(V1065+V1474)について、支払額が0の世帯はそのまま0、プラスの世帯は自然対数変換した値
貯蓄型保険		貯蓄型保険料の支払い額(V0613)について、支払額が0の世帯はそのまま0、プラスの世帯は自然対数変換した値
自営業ダミー	就業状況	世帯主/世帯2人目の企業区分が自営(V0045/V0059=2)の場合に1をとるダミー変数
大企業ダミー		世帯主/世帯2人目の企業区分が民営(V0045/V0059=1)かつ、企業規模が500人以上(V0046>=4)の場合に1をとるダミー変数
中小企業ダミー		世帯主/世帯2人目の企業区分が民営(V0045/V0059=1)かつ、企業規模が500人未満(V0046<=3)の場合に1をとるダミー変数
公務員ダミー		世帯主/世帯2人目の企業区分が官公(V0045/V0059=3)の場合に1をとるダミー変数
非就業ダミー		世帯主/世帯2人目の就業・非就業の別が非就業(V0044/V0058=3 or 4)の場合に1をとるダミー変数　※推定の際には、この変数を基準に設定する
経常所得	その他世帯属性	経常所得(V0402)を自然対数変換した値
貯蓄		貯蓄残高(V0671)について万円単位を円単位にしたものを自然対数変換した値
女性ダミー		世帯主性別；女性(V0393)が1の場合に1をとるダミー変数
世帯主年齢		世帯主年齢に関する5歳刻みの区分データ(V0042)
世帯人数		世帯人員数(V0017)
18歳未満人数		18歳未満の世帯人員数(V0388)
65歳以上人数		65歳以上の世帯人員数(V0389)
大都市圏ダミー		居住地域が3大都市圏(V0016=1)の場合に1をとるダミー変数

2-2. 基本統計量、散布図

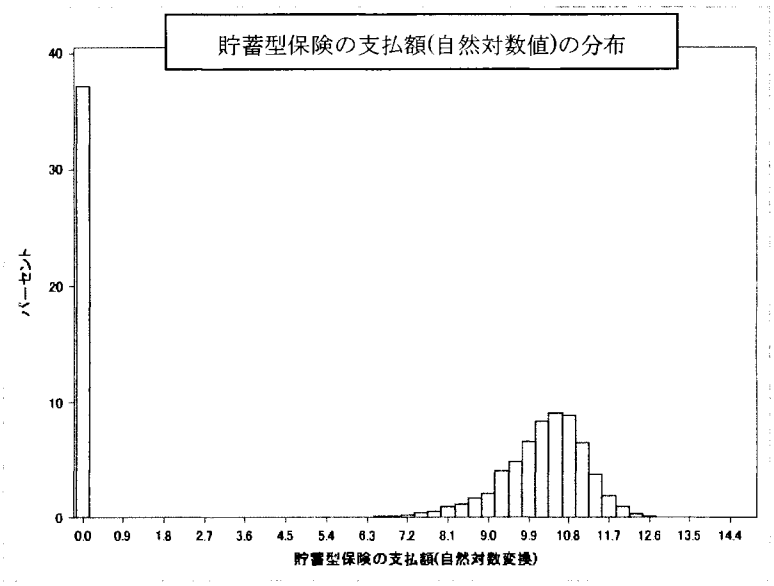
2-1 で設定した y_1:非貯蓄型保険料支払額(自然対数変換値)、y_2:貯蓄型保険料支払額(自然対数変換値)それぞれの基本統計量及び度数分布図を図 2-2-1、図 2-2-2 に示した。また、各保険の加入・非加入の 2 値のマトリクスを表 2-2-1 で示した。さらに、両保険料支払額が正の世帯について、散布図を図 2-2-3 に示した。これらに加え、本稿で着目する就業属性について、世帯主の就業属性で切り分けを行った両保険支払額の基本統計量を表 2-2-2 および表 2-2-3 に示した。

図 2-2-1　非貯蓄型保険料支払額の度数分布図、基本統計量



非貯蓄型保険	
N	43861
N(>0)	26289
mean	8.59
std	1.13
min	1.20
Q1	7.94
median	8.69
Q3	9.37
max	12.77

図 2-2-2　貯蓄型保険料支払額の度数分布図、基本統計量



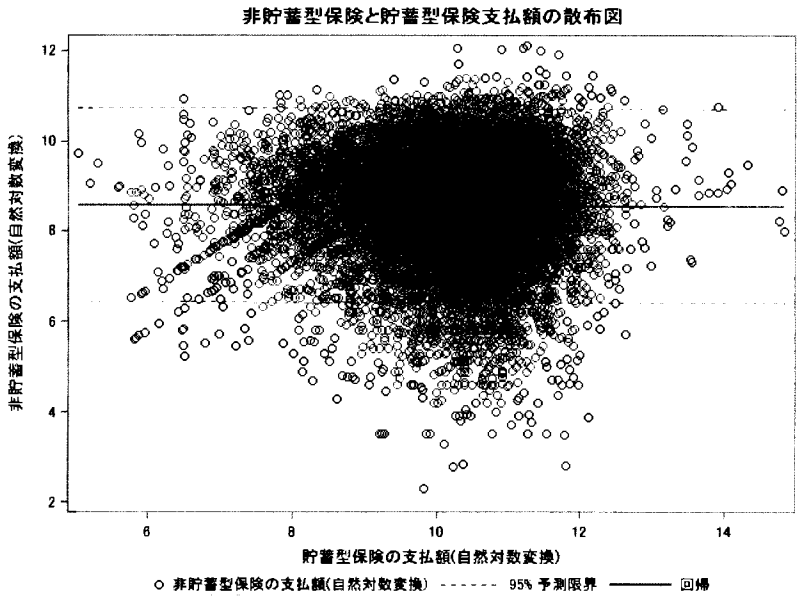
貯蓄型保険	
N	43861
N(>0)	27527
mean	10.23
std	0.96
min	5.04
Q1	9.71
median	10.35
Q3	10.87
max	14.83

表 2-2-1　各保険の加入・非加入世帯数の度数分布表

	貯蓄型：加入	貯蓄型：非加入	合計
非貯蓄型：加入	18906 世帯	7383 世帯	26289 世帯
非貯蓄型：非加入	8621 世帯	8951 世帯	17572 世帯
合計	27527 世帯	16334 世帯	43861 世帯

上記の表から、非貯蓄型・貯蓄型の両方に加入している世帯が 40%程度、非貯蓄型・貯蓄型のいずれか一方に加入している世帯がそれぞれ 20%程度、いずれの保険にも加入していない世帯が 20%となった。各保険の加入世帯について見ると、非貯蓄型保険では平均値が 8.59(指数変換すると約 5400 円)の比較的正規分布に近い分布となる。また、貯蓄型保険についても、平均値が 10.23(指数変換すると約 27700 円)の比較的正規分布に近い分布となることが分かった。

図 2-2-3　非貯蓄型保険と貯蓄型保険支払額の散布図(いずれの支出も正のサブサンプル)



以上の図から分かる通り、非貯蓄型と貯蓄型両方の保険に加入している世帯(18906 世帯)では、両方の保険の保険料支払額について、顕著な相関は確認できなかった。

さらに、本稿で着目する就業状況について考察するため、世帯主の就業属性で区分した両保険の基本統計量を以下の表 2-2-2 および表 2-2-3 で確認する。

表 2-2-2　世帯主の就業状況で区分した非貯蓄型保険料支払額の基本統計量

非貯蓄型保険	世帯数	加入世帯数	加入率(%)	平均値	標準偏差
非就業・その他	9855	4757	48.3	8.47	1.12
自営業	7060	3192	45.2	8.67	1.17
大企業	6479	4923	76.0	8.60	1.12
中小企業	15025	9280	61.8	8.60	1.11
公務員	5442	4137	76.0	8.59	1.15

表 2-2-3　世帯主の就業状況で区分した貯蓄型保険料支払額の基本統計量

貯蓄型保険	世帯数	加入世帯数	加入率(%)	平均値	標準偏差
非就業・その他	9855	6076	61.7	10.05	1.02
自営業	7060	6	0.1	9.88	0.35
大企業	6479	5567	85.9	10.26	0.92
中小企業	15025	11046	73.5	10.22	0.94
公務員	5442	4832	88.8	10.47	0.92

表 2-2-2 から分かるように、非貯蓄型保険に関しては就業状況に応じて加入率が異なり、非就業や自営業に比べ、大企業、中小企業、公務員の加入率が高いことが確認できた。また、加入世帯の平均値に着目すると、非就業に比べ他の世帯が若干高い傾向が確認できる。

次に、貯蓄型保険の表 2-2-3 について考察する。加入率に着目すると、自営業世帯が突出して低いことがわかる。また、非貯蓄型と同様に、非就業と比較して大企業、中小企業、公務員の世帯では加入率が高い傾向が確認できる。さらに、加入世帯の平均値についても、非就業と比較して大企業、中小企業、公務員の世帯では支払額が若干高い傾向が確認できる。

本節では分析で用いるデータセットの特徴を述べ、基本統計量等について確認した。次節では、上記で得られた両保険の加入行動の傾向を検証するための、推定モデルについて解説する。

3. 推定モデル

本節では、推定モデルの構成を解説し、最尤推定する尤度式を示す。本稿で分析に用いた推定モデルは、①非貯蓄型保険需要に関する推定式、②貯蓄型保険需要に関する推定式の 2 式で構成される。非貯蓄型保険、貯蓄型保険のいずれも負の値を取りえないことから、前節で確認したように支払額が 0 の点での打ち切り分布を形成している。このため、通常の線形回帰モデルを用いると、推定結果にバイアスが生じることが知られている。そのため、本稿では非貯蓄型、貯蓄型それぞれの加入選択と支払額選択を同時に扱える、SUR(Seemingly Unrelated Regression)Tobit モデルを用いて推定を行う。これは連立方程式推定の一種で、推定モデルの誤差分散構造のみに関係性を盛り込んだモデルである。負の値も取り得る潜在的な需要変数(y^*)を用いて、数学的には以下のように記述できる。

$$\begin{cases} y_1^* = \mathrm{X}\beta + \varepsilon_1 \\ y_1 = \max(0, y_1^*) \end{cases}$$
$$\begin{cases} y_2^* = \mathrm{X}\gamma + \varepsilon_2 \\ y_2 = \max(0, y_2^*) \end{cases}$$
$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} {\sigma_{11}}^2 & \sigma_{12} \\ \sigma_{12} & {\sigma_{22}}^2 \end{pmatrix} \right)$$

上式の通り、非貯蓄型保険、貯蓄型保険の両方について、潜在的な保険需要(y_1^*と y_2^*)の値に応じて、実際の保険料支払額が決定される状況が記述できる。2節で確認した通り、各保険の加入・非加入の区分が4つに分かれることから、それぞれの状況に対応した4通りの尤度関数を以下のように考えることができる。

$$if \quad y_1 = 0 \quad and \quad y_2 = 0 \quad then$$
$$L = \Phi_2\left(\frac{\mathrm{X}\beta}{\sigma_{11}}, \frac{\mathrm{X}\gamma}{\sigma_{22}}, \frac{\sigma_{12}}{\sigma_{11}\sigma_{22}} \right)$$
$$if \quad y_1 = 0 \quad and \quad y_2 > 0 \quad then$$
$$L = \frac{1}{\sigma_{22}} \phi\left(\frac{y_2 - \mathrm{X}\gamma}{\sigma_{22}} \right) \left[1 - \Phi\left(\frac{1}{\sqrt{1 - (\sigma_{12}/\sigma_{11}\sigma_{22})^2}} \left(\frac{\mathrm{X}\beta}{\sigma_{11}} + \frac{\sigma_{12}}{\sigma_{11}\sigma_{22}} \frac{y_2 - \mathrm{X}\gamma}{\sigma_{22}} \right) \right) \right]$$
$$if \quad y_1 > 0 \quad and \quad y_2 = 0 \quad then$$
$$L = \frac{1}{\sigma_{11}} \phi\left(\frac{y_1 - \mathrm{X}\beta}{\sigma_{11}} \right) \left[1 - \Phi\left(\frac{1}{\sqrt{1 - (\sigma_{12}/\sigma_{11}\sigma_{22})^2}} \left(\frac{\mathrm{X}\gamma}{\sigma_{22}} + \frac{\sigma_{12}}{\sigma_{11}\sigma_{22}} \frac{y_1 - \mathrm{X}\beta}{\sigma_{11}} \right) \right) \right]$$
$$if \quad y_1 > 0 \quad and \quad y_2 > 0 \quad then$$
$$L = \frac{1}{\sqrt{{\sigma_{11}}^2 {\sigma_{22}}^2 - {\sigma_{12}}^2}} \phi_2\left(\frac{y_1 - \mathrm{X}\beta}{\sigma_{11}}, \frac{y_2 - \mathrm{X}\gamma}{\sigma_{22}}, \frac{\sigma_{12}}{\sigma_{11}\sigma_{22}} \right)$$

式中のΦ、ϕはそれぞれ1変量標準正規分布の累積分布関数、確率密度関数を表す。また、Φ_2、ϕ_2はそれぞれ2変量標準正規分布の標準正規分布の累積分布関数、確率密度関数を表す。

次に、モデルを構成する説明変数の設定について述べる。両保険の支払額を考察するために、それぞれに影響を与えると考えられる世帯属性として、以下のものを説明変数に設定した。

・世帯主の就業属性(非就業・その他世帯を基準として、自営業、大企業、中小企業、公務員)
・世帯2人目の就業属性(非就業・その他世帯を基準として、自営業、大企業、中小企業、公務員)
・その他世帯主属性(世帯主女性ダミー、世帯主年齢区分)
・世帯属性(世帯人員数、18歳未満人員数、65歳以上人員数、大都市圏ダミー)
・収入等(経常所得対数値、貯蓄対数値)

以上の設定をもとに、上述の尤度関数が構成でき、この式をもとにパラメータの最尤推定値を求める。推定に先立ちパラメータの初期値を指定する必要がある。これについては、各推定式単体でTobitモデルの最尤推定を行って得たパラメータ推定値を、初期値として設定した。

推定プログラムの概要は、以下の通りである。詳細については、別途プログラム本体を参照されたい。

```
proc nlmixed data=data tech = NEWRAP;
 /*パラメータの初期値設定*/
 parms b100 ～ b230 s11 ～ s22;   bounds s11 s22 > 0; pi=atan(1)*4;
 y1 = log_kakesute; y2 = log_V0613;

 /*右辺を定義する*/
 xbeta = b100 + b101*Jiei1 +… + b130*log_saving ;   xgamma =   b200 + b201*Jiei1 +… + b230*log_saving ;

 /*誘導系のモデル式を定義する*/
 e1 = y1 - xbeta;   e2 = y2 - xgamma;

 /*分散共分散行列の行列式を定義する*/
 det_sig = s11**2*s22**2 - s12**2;

/*尤度関数を定義*/
 if y1 = 0 and y2 = 0 then
 ll = log(probbnrm(xbeta/s11,xgamma/s22,s12/(s11*s22)));

else if y1 = 0 and y2 > 0 then
 ll = -log(s22) + log(PDF('NORMAL',e2/s22, 0,1))
 + log(1-CDF('NORMAL',1/(1-(s12/(s11*s22))**2)**0.5
 * (xbeta/s11+(s12/(s11*s22))*e2/s22),0,1));

else if y1 > 0 and y2 = 0 then
 ll = -log(s11) + log(PDF('NORMAL',e1/s11, 0,1))
 + log(1-CDF('NORMAL',1/(1-(s12/(s11*s22))**2)**0.5
 * (xgamma/s22+(s12/(s11*s22))*e1/s11),0,1));

else if y1 > 0 and y2 > 0 then
 ll = -1/2*log(2*pi) -1/2*log(det_sig) -1/det_sig
        *(s22*e1**2 - 2*s12*e1*e2 + s11*e2**2) ;

model log_V0613 ~ general(ll);
run; /*尤度関数の最大化を定義*/
```

4. 推定結果および考察

3節で述べた推定プログラムを実行することにより、以下の結果を得ることが出来る(表4-1)。一般的な線形モデルでは、対数変換した被説明変数に対して、説明変数の推定値はカテゴリー変数であれば基準に対する比、連続量であれば変化率や弾力性として解釈できる。しかし、Tobitモデルでは加入・非加入の2値選択と、支払額の量的選択の傾向を同一パラメータで推定しているため、線形モデルのような解釈を行うことができない点に注意が必要である。以下では各世帯属性について推定結果を概観し、簡単に解釈を行う。

表4-1　推定結果

	非貯蓄型保険需要の推定			貯蓄型保険需要の推定		
	推定値	標準誤差	p値	推定値	標準誤差	p値
世帯主自営業ダミー	2.238	0.146	<.0001	-3.350	0.113	<.0001
世帯主大企業ダミー	1.225	0.089	<.0001	0.305	0.070	<.0001
世帯主中小企業ダミー	0.502	0.079	<.0001	-0.132	0.062	0.033
世帯主公務員ダミー	1.224	0.091	<.0001	0.485	0.071	<.0001
2人目自営業ダミー	-0.165	0.118	0.163	0.369	0.093	<.0001
2人目大企業ダミー	0.326	0.095	0.001	0.046	0.075	0.538
2人目中小企業ダミー	0.169	0.054	0.002	0.114	0.042	0.007
2人目公務員ダミー	0.447	0.085	<.0001	0.380	0.066	<.0001
世帯主年齢	0.050	0.013	<.0001	0.046	0.010	<.0001
世帯人員数	0.127	0.027	<.0001	0.173	0.021	<.0001
18歳未満人数	-0.138	0.035	<.0001	0.034	0.027	0.207
65歳以上人数	-0.380	0.038	<.0001	-0.021	0.029	0.483
大都市圏ダミー	0.100	0.044	0.023	-0.387	0.035	<.0001
世帯主女性ダミー	-0.582	0.085	<.0001	-0.273	0.067	<.0001
経常所得対数値	-0.027	0.010	0.005	0.433	0.008	<.0001
貯蓄対数値	0.018	0.005	0.000	0.022	0.004	<.0001
	N of observation			43861		
	-2* Log Likelihood			323635		
	AIC			323709		

【世帯主の就業状況について】

非貯蓄型保険では、無職世帯を基準としていずれの就業状況も高い保険需要となり、特に自営業世帯が突出して支払額が高い傾向が確認できた。一方で、貯蓄型保険では、逆に自営業の保険需要が著しく低い傾向が確認できた。また、全体的に公務員世帯は非貯蓄型、貯蓄型ともに高い保険需要を示す傾向が確認できた。こうした傾向は基本統計量ベースで検討した2節の結果と整合的である。

自営業世帯の傾向が突出して見られた点については、貯蓄型保険に関する何らかの制度(例えば、自身で運営している事業の方で計上するなど)が影響している可能性などが考えられる。濵本(2001)が指摘している通り、自営業世帯では世帯主に万一の事態があった場合に、一般従業世帯よりも影響が大きいことから保険需要は高まると考えられる。この需要を非貯蓄型の保険でカバーしている可能性も否定できないが、やはり何らかの制度上の影響があると思わる。この点についての検証は今後の課題としたい。

【世帯2人目の就業状況について】

全体的に、世帯主の就業状況と比較すると影響の度合いは軽微であった。若干の傾向として、非貯蓄型、貯蓄型ともに公務員世帯では支払額が高い傾向が見られ、公務員共働き世帯で高い保険需要となることが確認できた。また、世帯主の就業状況とは逆に、2人目が自営業の世帯では貯蓄型の需要がやや高まる傾向が確認できた。

【世帯主属性について】

世帯主の年齢が高まるにつれ保険料の支払額は高まり、その傾向は非貯蓄型と貯蓄型で同程度であることが確認できた。一般的に高齢になるほどライフサイクル動機での保険需要は低下すると考えられるため、今回の推定結果は遺産動機の発現、もしくは年齢の高まりにより保険料の設定自体が高まることが原因と解釈できる。また、世帯主が女性の世帯では、非貯蓄・貯蓄ともに低い水準であった。

【世帯属性について】

世帯人数の増加に伴い、保険料の支払額も高まる自明の結果が確認できた。また、子供が多い世帯では非貯蓄型よりも貯蓄型を選択する傾向が強いことが確認できた。一方で、高齢者が多い世帯では非貯蓄型の支払いは少なくなるものの、貯蓄型については顕著な傾向が見られなかった。また、大都市圏では貯蓄型保険の需要が低い傾向が確認できた。

【所得・貯蓄について】

高所得世帯では非貯蓄型よりも貯蓄型を選好する傾向が確認できた。また、貯蓄額については非貯蓄型に対しては影響しないものの、貯蓄型に対しては有意な正の影響が確認でき、世帯の資産選好と整合的な結果が得られた。

5. まとめと今後の課題

本稿では平成16年度の総務省全国消費実態調査匿名データを用い、非貯蓄型保険と貯蓄型保険両方の需要に対して、世帯属性の中でも特に就業状況が与える影響に着目した。推定モデルには、連立方程式の誤差の分散構造のみに関係性を盛り込んだ SUR(Seemingly Unrelated Regression) Tobit を用いた。

推定の結果、就業状況が保険需要に与える影響は、非貯蓄型と貯蓄型で大きく異なることが確認できた。特に、自営業世帯では非貯蓄型保険需要が突出して高い一方で、貯蓄型保険需要が突出して低い特徴的な傾向が確認できた。また世帯2人目までの就業状況を考慮すると、公務員世帯が他の世帯と比較して非貯蓄型、貯蓄型のいずれも高い需要となることが確認できた。

次に、課題と今後の展望について述べる。今回の分析で自営業世帯の保険加入行動の独自性が明らかとなったが、その原因としていくつかの可能性を提示したものの、詳細を明らかにすることは困難である。そのため、こうした需要の独自性が生じる原因については今後検討を行いたい。また今回の分析では SUR Tobit モデルで推定を行ったが、この推定方法には①同時決定モデルではないことから、保険加入選択の同時性を厳密には考慮できていない点、②加入・非加入の 2 値選択と支払額の量的選択の傾向を同一パラメータで推定している点、などで問題がある。さらに推定の都合上、世帯主年齢や世帯人員数など本来はカテゴリー変数として扱うものについて、単調性を仮定した連続量として推定している。こうした問題点をクリアするための方法として、2 値選択と量的選択の傾向を別のパラメータで推定する Hurdle モデル(Two-part モデル)を同時方程式モデルに拡張する方法が考えられる。一段と複雑な尤度関数を推定することになるため、今後の課題としたい。

6. 参考・引用文献

・Wooldridge, J., (2010), Econometric Analysis of Cross Section and Panel Data (2nd ed.), MIT Press

・浅野 晢, (1998), 「公的年金制度と個人年金、生命保険需要」, 日本経済研究 No.36, p.83-102

・岩本 光一郎, (2003), 「保険需要の要因分析：家計のライフサイクルの視点から」, 「生命保険に関する全国実態調査」の再分析, 第 2 章

・濵本 浩幸, (2001), 「生命保険金額に影響を及ぼしている原因」, 『郵政研究所月報』2001.2

ビジネス活用

SASユーザ総会 2014発表資料

ビジネスにおけるビッグデータ活用の歴史と今後の展望

坂巻　英一
公立大学法人宮城大学
事業構想学部

History and Future Prospect of Big Data in Business Area

Yoshikazu, SAKAMAKI
Department of Project and Design

要旨

近年，メディア等でビッグデータという言葉を耳にする機会が多い。企業活動の仮定に於いて蓄積されたテラバイト，ペタバイトといった市販されている分析ツールでは分析することが困難な巨大なデータを指す。近年，こうしたデータを効率的に処理する技術が次第に確立されつつあり，企業の経営効率改善に役立てようとする試みも盛んになってきている。巨大なデータを処理するに当たり必要となるコンピュータの性能向上と共に，RDB を中心としたデータベースだけではなく，並列分散処理を実現するソフトウェアとして Hadoop が無料で使用できるようになったことが，ビッグデータの活用を後押ししていると言えよう。

ところが，企業に蓄積された大規模データを分析し，そこから得られた知見を企業の経営効率改善に役立てようという動きは 1990 年代の前半には既に始まっていたのである。有名な事例として POS データを分析しそこから得られた知見を基に店舗のレイアウトを変更したところ，売上改善を実現することができた，という報告が挙げられる。こうしてみると，ビッグデータやデータサイエンスに関連する技術は今に始まったものではないことに気付かされる。本稿ではビジネスにおけるビッグデータ利用の歴史について概観すると共に，ビッグデータ活用の今後の展望について概観する。

キーワード：ビッグデータ，データサイエンス，データマイニング

1. ビールと紙おむつの事例

論文要旨で既述したように，ビッグデータを分析した結果得られた知見をビジネスで活用した事例は1990年代の前半にまで遡る。有名な事例として，アメリカのあるスーパーマーケットチェーンが行ったPOSデータの分析事例が有名であろう。これは店舗に蓄積されたPOSデータを分析した結果，金曜日の夜になると多くの顧客が缶ビールと紙おむつを一緒に購入する傾向があることが分かった，というものである。この店舗では分析結果を基に，缶ビール売り場の横に紙おむつを並べたところ，缶ビールと紙おむつの売上が共に上がった，というのだ。分析結果は1992年12月に発行されたWall Street Journalで紹介され，ビッグデータ分析の初期の頃の代表的な成功事例として後に広く知られるようになった。

こうした分析手法は一般に併買分析と呼ばれており，現在，流通小売業界において最もよく行われる分析手法のひとつである。POSデータを分析することで得られた知見を基に，同時に購買される傾向のある商品を見つけ出し，陳列棚の棚割りを決定する際に利用されている。併買分析はリアル店舗だけではなく，ネットショップでも利用しているサイトは多い。Amazon.com等で商品を購入した際に，他の商品を勧められたことはないだろうか。これはレコメンデーションエンジンと呼ばれており，併買分析の結果を活用した典型的な事例であると言える。

2. データサイエンティストという仕事

ビッグデータを分析する技術者は，最近，データサイエンティストと呼ばれており，近い将来，人材不足に陥る可能性が高いと言われている。ところが，データサイエンスという言葉自体，既に様々な場所で使われているにも拘らず，データサイエンティストが行う仕事については，未だに定義があいまいなままである。最近の求人広告を見ていると，データサイエンティストと呼ばれる職種は大きく分けて2種類に分かれているように思われる。

一つ目がビッグデータの管理等を効率的に行う，SEタイプの職種である。Hadoopのコーディングや Javascriptを利用した帳票を出力させるためのWEB画面の構築もここに含まれる。そして二つ目がアナリストタイプの職種である。営業部門やマーケティング部門と密に連絡をとりながら，データベースから抽出されたデータを解析し，得られた知見を営業やマーケティングの現場へフィードバックする需要な役割を果たすことになる。データサイエンティスト募集，と書かれた求人広告の中でよく目にするSASやSPSS，R等のBIツールのオペレーションスキルを持った人材がここに該当する。

ところで，SASやSPSSを使用した分析業務はいつ頃からあるのだろうか。少なくとも，90年代の中ごろには，企業が保有する顧客データを分析し，得られた知識を経営改善に生かそうという取り組みが存在していたように記憶している。こうして見ても，データサイエンティストと呼ばれる職業が今に始まったものではないことに気付かされるのではなかろうか。

3. ビッグデータの活用を進めるために

「ビッグデータ」や「データサイエンス」という言葉が消滅しても，企業に蓄積された大規模データを活用しよう，という動きは引き続き残ると考えられる。なぜならば，大規模データのビジネスでの活用は今に始まったものではなく，20年近くも前から存在するからである。それでは，ビッグデータの活用を成功に導

くためにはどうしたらよいのだろうか。図１は社内にデータサイエンス専門の部隊を有するか否か，社内に独自の大規模データを保有しているか否か，を基に企業を４つのパターンに分類した結果である。

独自の大規模データを保有しており社内にデータサイエンティスト集団を抱えている企業は，既に，ビッグデータの活用が相当進んでいる企業である。こうした企業は社内にも十分なノウハウが蓄積されており，ビッグデータを日常的に活用した経営改善に取り組んでいると考えられる。問題は社内に独自の大規模データを保有しているにも関わらず，専門の部隊を有しない企業である。データサイエンスに関する技術やノウハウは一朝一夕では蓄積することができないため，外部のコンサルティング会社等に分析業務を依頼しなければならないことが多い。

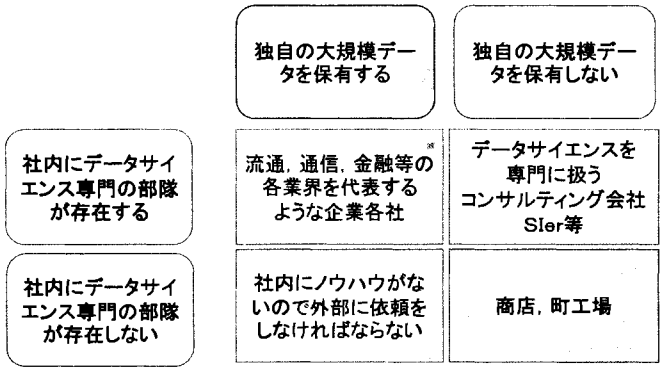


図１　保有データの規模ごとに見たデータサイエンスへの取り組み

ところが，ここで是非次のことを記憶に留めておいて欲しい。データサイエンスに関する技術やノウハウは技術者やアナリストの中に蓄積されるものだ。そのため，業務を外部に丸投げするようなプロジェクトの進め方をしたのでは，時間が経っても社内には全くノウハウが蓄積されない，という事態が発生しかねない。こうした方法でプロジェクトを進めていくと，気付いた時には社内のシステムがブラックボックス化し社内だけではデータサイエンスに関するプロジェクトを回せなくなっていることもあり得るのだ。やむを得ずシステム開発や分析業務を外部に委託する場合には，何を最終ゴールとするのか，プロジェクトの目的を経営陣が明確にしておく必要があると言える。

ビッグデータの活用を社内で成功させるためには，３つのＳが重要な役割を果たすことに気付かされる。それは，System(Database，Hadoop 等)，Statistics(統計解析)そして Strategy(経営戦略)である。これら３つの要素を全て兼ね備えた人材は地球上探し回っても数えるほどしかいない。つまり，これらの要素のうちのいずれかに秀でた人材を集め，データサイエンスチームを社内に構築した上でプロジェクトを推進してゆくことが求められるのである。

どんなに高価な情報システムや統計ソフトを導入しても，戦略が揺らいでいてはよい成果は得られない。また，どんなに良い戦略を構築しても，数値的な裏付けが十分でなければよい成果は得られない。これら３つのＳが融合した時に初めて，ビッグデータの活用は成功するのである。これを聞いて，初期の頃のデータマイニングを思い出す人がいるのではなかろうか。データマイニングがはやり始めた 90 年代の後半，データマイニングプロジェクトを成功に導くために必要な要素として，全く同じことが言われていたことに気付かされよう。

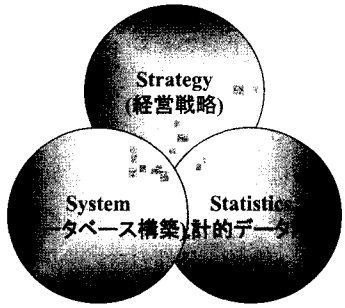


図2　データサイエンスに必要な要素

4. バズワードはいずれ消滅する

企業に蓄積された大規模データを分析し，企業の経営改善に活用する取り組みが現在様々な企業に於いて行われている。一方で，第3節で既述した通り労働市場に於いてデータサイエンティストと呼ばれる人材はそれほど多くないのが現状であり，今後，人材が不足することが予想される。一方で，起業に蓄積された大規模データを分析しデータの背後に潜む規則性や法則性を見つけ出す技術はかつて「データマイニング」と呼ばれており，20年以上前から存在するのである。「データサイエンティスト」や「ビッグデータ」という言葉は大手のメディアが騒ぎ立てたがゆえに，現在，バズワードになっているのではなかろうか。

これらの言葉の定義が未だに明確ではないことや，バズワードは必ず消滅する運命をたどってきた歴史を考えると，近い将来，「データサイエンティスト」や「ビッグデータ」という言葉は消滅するのではないかと考えられる。ところが，大規模データのビジネスでの活用は20年以上も前から世の中に存在していたのである。「データマイニング」から「ビジネスインテリジェンス」そして「ビッグデータ」「データサイエンス」，時間の流れと共に呼び方が変わっただけで，やっていることは20年以上前から何ら変わらないのだ。

5. まとめ

ビッグデータという言葉を耳にするようになってかなりの時間が経ったような気がする。

分析ツールは複雑な計算をアイコン操作や短いプログラムで実行することを可能にする。つまり，与えられた問題の答えを見つけ出すのは非常に得意である。ところが，分析ツールは何が問題なのか，までは考えてくれないのである。それを考えるのは社内の人間に他ならない。分析ツールは複雑な計算を行うための道具でしかなく，決して分析ツールを導入しただけでは，結果を出すことはできないのだ。SAS システムを始めとした分析ツールに加え，それをどのように活用したらよいか，を考える立場の人材がビッグデータを活用する上で必要なのだということを肝に銘じておく必要があると言えよう。

以上

DeloitteのAudit AnalyticsにおけるSAS Visual Analyticsの活用事例紹介

戸田大介
有限責任監査法人トーマツ/Deloitte Analytics/ジュニアスタッフ

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

当監査法人で取り組んでいるAudit Analyticsという監査業務を差別化するための分析サービスを紹介させて頂きます。

当法人のAudit Analyticsは会計監査の品質向上・効率化を実現しただけでなく、監査クライアントへ会計監査を通じて新たな知見を提供することをも可能としました。視覚的に財務データ・非財務データを識別することで監査チームは効率的にリスクエリアを特定し、監査を行うことが可能となり、また、大規模な監査対象の多岐にわたるデータを網羅的に集計・認識することで、漏れの少ない品質の高い監査を行うことが可能となりました。加えて、これらの分析結果を視覚的に監査報告書としてクライアントへ提供することで、会計監査を通じてクライアントビジネスへの気づきを提供することも可能となりました。
DeloitteのAudit AnalyticsではSAS Visual Analyticsを用いて分析を行っております。当該製品を用いたAudit Analyticsの分析事例をご紹介させて頂きたいと思います。

CDISC標準対応でSASプログラマーが抱える問題点と解決策

片山 雅仁 小山 卓己 山本 松雄
イーピーエス株式会社 CRO事業本部 DSセンター

Problems and solutions that SAS programmers faced with CDISC Standards.

Masahito Katayama, Takumi Koyama, Matsuo Yamamoto
Data Science Center, EPS Corporation

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

CDISC標準対応にあたり、SASプログラマーだけでは解決できない問題点がある。弊社では社内研修を通じてその解決策を展開している。

キーワード:CDISC SDTM ADaM CDASH SASプログラマー
業務プロセス 組織横断 教育研修

背景

1997年 Clinical Data International Standards Consortium (CDISC)発足

2004年 SDTM V1.0 リリース

～ FDAがCDISCによるデータ申請受付を開始～

～外資系製薬メーカーを中心にCDISCが普及～

2013年9月 PMDA「次世代審査・相談体制に関する説明会」

～製薬メーカー・CRO各社対応に追われる～

2016年 CDISC標準準拠電子データ提出の義務化開始

CDISC標準への各社対応

製薬メーカーA社

「うちは本国で導入するようになったので大丈夫！！」

製薬メーカーB社

「準備してきたけどCDISC Likeだから申請に耐えられるか自信がない。」

製薬メーカーC社

「2016年までに対応しなきゃいけない！！どうしよう…」

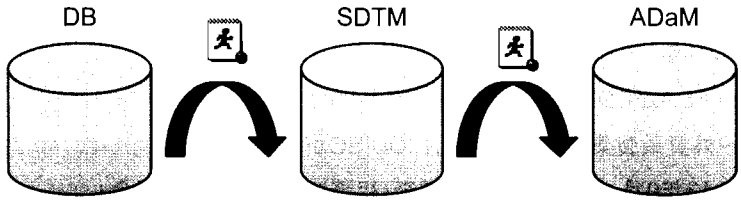
CRO X社

「これからどんどんCDISC案件が増えそうだ」

CRO Y社

「SDTMは誰が作ったらいいの？？SASは必要なの？？」

CDISC標準へのSASプログラマーの関わり



解析用のSASデータセットを作成する従来の業務に近いため
SDTM、ADaMを作成にSASプログラマーは重要な役割を果たしている。

SASプログラマーが直面した問題

SASプログラマー Aさん
「そもそもCRF、オリジナルのDBがCDISC標準からとても遠い…」

例）よくありがちな患者背景データのオリジナルDB

CRFNO	SEX	GEN	GAPPEIUM	KIOUUM
症例番号	性別	原疾患	合併症	既往歴
X001-01	1	2	1	1
X003-01	2	1	2	2

SDTMではMHドメインなのに

SASプログラマーが直面した問題

SASプログラマー Bさん

「測定項目やコードリストの区分がうまくSDTMにマッピングできない…」

例）併用薬剤のCRF

併用薬剤名：＿＿＿＿＿＿
投与経路
☐ 1:経口
☐ 2:注射
☐ 3:外用
☐ 9:静注

このTerminologyだけでも
100個以上ある！！
経口でも色々ありすぎる！！

NCI Controlled Terminology

Code 66729

ROUTE (Route of Administration)

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

SASプログラマーが直面した問題

SASプログラマー Bさん（つづき）

「測定項目やコードリストの区分がうまくSDTMにマッピングできない…」

例）有害事象の転帰のCRF

CRFとTerminologyが一致しない！
しかもTerminologyは変更不可

☐ 1:回復
☐ 2:軽快
☐ 3:後遺症あり
☐ 4:未回復
☐ 5:死亡
☐ 6:死亡未回復
☐ 7:不明

NCI Controlled Terminology

FATAL
NOT RECOVERED/NOT RESOLVED
RECOVERED/RESOLVED
RECOVERED/RESOLVED WITH SEQUELAE
RECOVERING/RESOLVING
UNKNOWN

SASプログラマーが直面した問題

■問題点を整理すると...

- ➢SDTMやADaMはSASプログラマーが作成するしかない！
- ➢CRF、オリジナルのDBがSDTMからかけ離れていて苦労する。

⇒そもそもSASプログラマーが全て考えるべきことでしょうか？

⇒SASプログラマーがCDISC標準を学習すれば対応できるか？
⇒DB設計者がCDISC標準を学習すれば対応できるか？
⇒CRFは？プロトコル設計者は？

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

解決策はあるのか?

■誰がCDISC標準を学習すべきなのか

- ➢Clinical Programmer ?
- ➢Data Manager ?
- ➢Bio Statistician ?
- ➢Medical Writer ?
- ➢薬事 ?
- ➢CRA ?
- ➢非臨床 ?
- ➢Medical Affairs ?

⇒臨床データのLifecycleに関わる全ての人が学習するのが理想

解決策はあるのか?

■どうやってCDISC標準を学習すべきなのか

- ➢CDISC Official Trainingの受講
- ➢学会(DIA、SASユーザー会など)に参加
- ➢CJUG、phUSEなどのユーザーグループに参加
- ➢SNS(LinkedInなど)のユーザーグループに参加
- ➢Webinerに参加
- ➢個人でImplementation Guideを読む
- ➢社内にCDISCプロジェクトチームを作る

⇒人、物、金、時間、情報を厭わなければ選択肢は豊富

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

解決策はあるのか?

■いつまでにCDISC標準を学習すべきなのか

- ➢2016年度(+移行期間2年程度)のCDISC標準の義務化まで
- ➢有識者の確保、組織の見直し等の準備期間は必要

⇒今から始めれば、まだ間に合う

⇒弊社では、2012年4月から準備を始めました

2012年4月　社内CDISCプロジェクト発足

■CDISC標準対応の推進

■当初はSASプログラマーが直面する問題解決のテーマにしていたわけではない

■しかし、プロジェクトの活動を通じて解決の手がかりが見えてきた

プロジェクト初期の活動内容

■DM、統計解析、システム開発と組織横断的な体制

■CDISC標準の勉強会実施

- IG、Controlled Terminology、TA Standrds、CDISC Implementation using SASなど。。。

■業務フローの検討、自社標準テンプレートの作成

■主にプロジェクト内部の活動だった。準備期間。

■社内に展開する必要がある。

社内CDISC研修の実施

■カリキュラム紹介
- ➢CDISC全般、SDTM、CDASH、ADaM
- ➢IG、Controlled Terminology、Define-XML、Reviewer's guide。。
- ➢ディスカッション、事例紹介、お悩み相談

■第1期、第2期

■各期約3ヶ月

■卒業テストあり

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

社内CDISC研修のコンセプト

■CDISC標準の中核はSDTM
- ➢SDTMからCDASHへ、SDTMからADaMへ

■業務担当者間の連携が不可欠
- ➢共通認識
- ➢業務プロセスの確認
  - ➢CDASH⇒SDTM⇒ADaM⇒TLF⇒CSR
  - ➢担当業務以外への理解

社内CDISC研修で伝えたかったこと

■CDISC標準に準拠した新薬申請パッケージ

➢SDTM、ADaM、TLF。付随するMetadata

➢SASプログラマーが申請資料作成に関わる機会は多い

⇒SASプログラマーだけで対応できるか？

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

社内CDISC研修で伝えたかったこと

■臨床データに携わるすべての担当者が知るべき

➢なぜ？

⇒SDTMは高度に標準化されている。
円滑に作成するにはそれなりの準備と仕組みが必要。

➢CRF設計、DB設計⇒CDASH

➢プロトコル

➢治験実施スケジュール

➢評価項目

社内CDISC研修の効果

■まずは知ってもらう。気付いてもらう。

- ➢SDTMが中核
  - ➢SDTM作成を見据えたCDASH
- ➢個人の知識、スキルも大切。でもそれだけではない。
- ➢ましてやSASプログラマーだけでは対応できないことも多々ある。

社内CDISC研修受講者の感想（DM担当者）

■立上げの段階の意識を変える

■DB設計の教育

■CRF設計からCDISCを意識しなければならない

■DM内での業務分担

■関連部署との連携強化が必要

■他の試験を意識する必要がある

■システム担当者でもCDISCの知識が必要

■CDASHだけではなく全体の流れを把握する必要あり

■上流工程で工数がかさんでも下流工程で削減できる

など

今後の展開

■もっと知ってもらう。

➢CRA、非臨床などのLifecycleに関わる人への研修の展開

プラットフォーム

アマゾン ウェブ サービス(AWS)による公共データの活用

吉荒 祐一

アマゾン データ サービス ジャパン株式会社

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

公共機関によるオープンデータの推進には、利用しやすい形でデータを公開する事と、データの解析によりビジネスの価値を生む事の二つの重要な要素があります。アマゾン ウェブ サービスが、いかに、このオープンデータの両輪を力強く推進しているかご紹介します。

SAS Loves Big Data via Hadoop
～Big Data Driven Innovation～

惟高 裕一，北西 由武，都地 昭夫

塩野義製薬株式会社

SAS Loves Big Data via Hadoop
～Big Data Driven Innovation～

Yuichi Koretaka , Yoshitake Kitanishi , Akio Tsuji

SHIONOGI & CO., LTD.

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨：

シオノギで構築したHadoop環境の紹介，HadoopとSASを連携させる方法，およびその留意点，さらにはそれらを利用したデータ解析事例について報告する．

キーワード：Big Data，Hadoop，hive，HDFS，Open Data，
SAS/ACCESS Interface to Hadoop

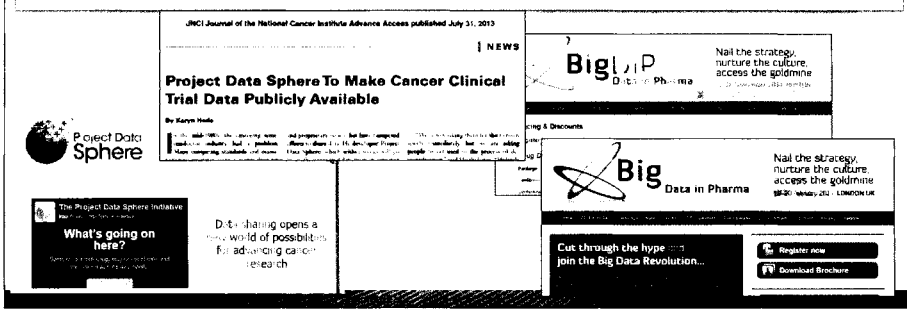
内容

- 背景
- Hadoop環境の紹介
  - Hadoopとは
  - シオノギのHadoop環境
- SASとHadoopの連携
  - SAS/ACCESS Interface to Hadoopに触れてみて
  - 解析事例

Big Dataの近況(特に医薬関連)

世界では,

- 各製薬会社がデータを提供し相互利用化が進み始めている(Data Sphere)
- EUでは透明性を主目的にした臨床試験データ公開の動きがある
- 医薬品産業会でのビッグデータ活用に向けて新たな学会が作られてきている (Big DIPなど)



医薬関連の Big Data (1/2)

- 製薬業界におけるBig Data解析ニーズの高まり
  - 社内データ
    - PGx
    - Safety Surveillance
  - オープンデータ
    - 添付文書データベース
    - 医薬品副作用データベース (JADER/AERS)
    - Real World Data (Claim data, EHR)
    - Open FDA
    - その他多数

* EHR: Electronic Health Record

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

医薬関連の Big Data (2/2)

新しい知見が得られるかもしれない

医薬品副作用DB　シグナル検出　添付文書DB

Claim data　市場調査　EHR

アイデア次第でDBの組み合わせは多数考えられ，組み合わせ数に応じてデータは大きくなってくる

ひとつの選択肢として

データ量と処理リソースの膨張し，
処理速度やデータ容量がネックとなってくる

Hadoop環境での解析
- 高度なJavaプログラミングの知識が必要
- 製薬会社のプログラマにとっては敷居が高い

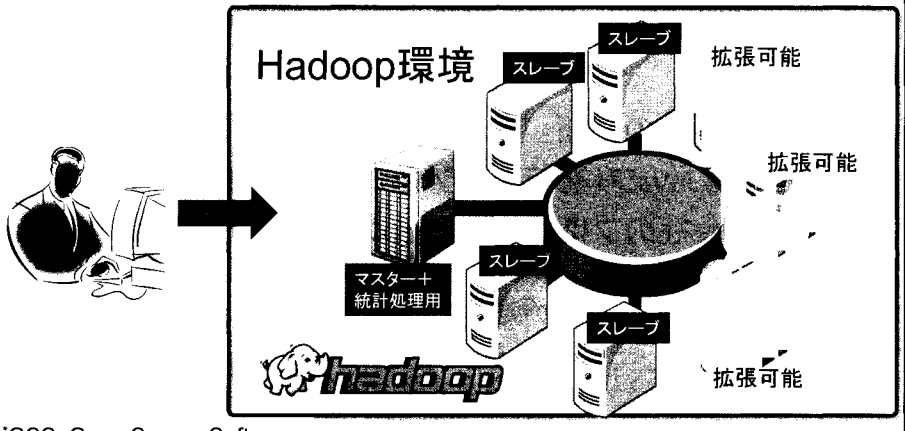
SAS/ACCESS Interface to Hadoopの利用を考えた

内容

- 背景
- Hadoop環境の紹介
  - Hadoopとは
  - シオノギのHadoop環境
- SASとHadoopの連携
  - SAS/ACCESS Interface to Hadoopに触れてみて
  - 解析事例

Hadoopとは

Googleの基盤技術に基づき，OSS*として実装された大規模分散処理フレームワーク

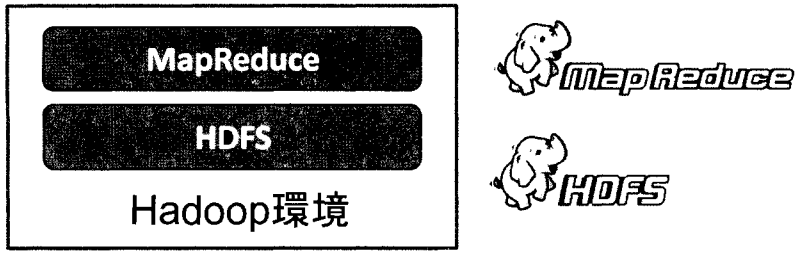


*OSS: Open Source Software

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

Hadoopとは

- HDFS*という分散ファイルシステム，MapReduceという分散処理システムを基本機能とする

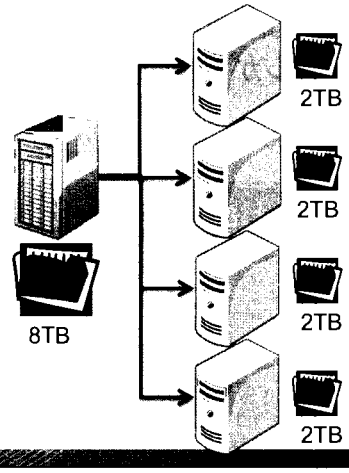


*HDFS: Hadoop Distributed File System

HDFS

分散ファイルシステム➔1つのファイルを分散して保持する

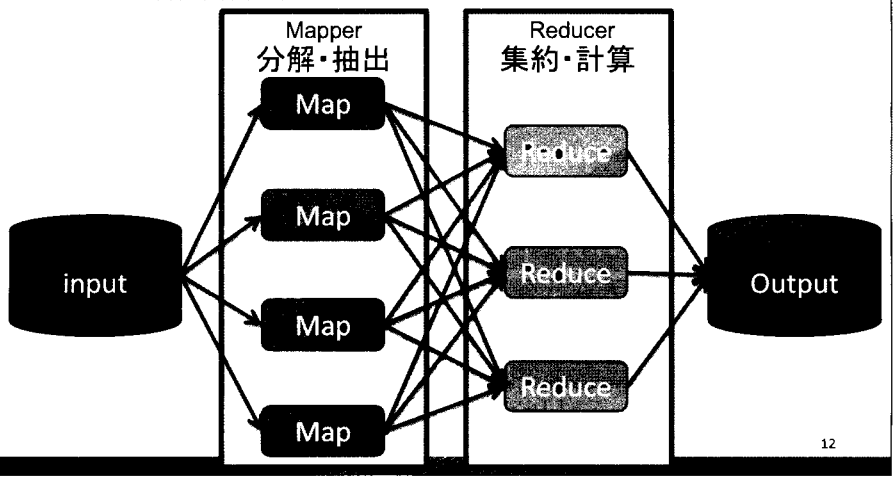
- 1台のPCでは扱えないようなサイズのデータを扱える
- 実際は分割したデータのコピーも保存されており，どこかのPCが壊れても問題ない



医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

MapReduce

分散処理システム➔処理を分散させて行う



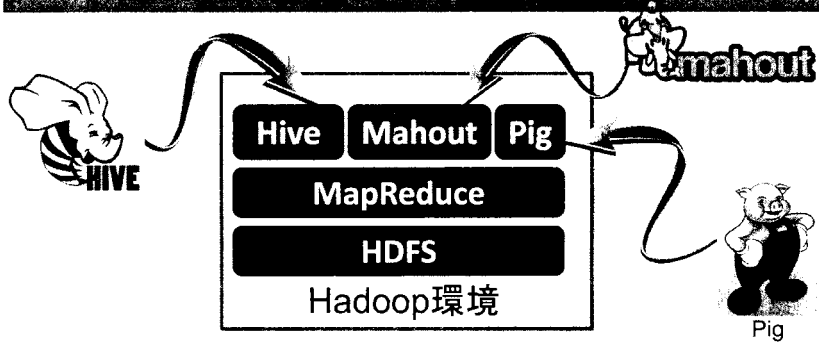
Hadoop (HDFS+MapReduce)

- 基本のHadoop環境だけで出来る処理は限られる
- MapReduce処理のためには，Javaのプログラミングスキルが必要であり，敷居が高い

Hadoopエコシステムが提供されている

Hadoopエコシステム

- Hadoopの使いにくい面を補うものとして，様々なエコシステムが提供されている



Hive

- HiveQLというSQLライクな言語でHDFS上に存在するデータを操作できる

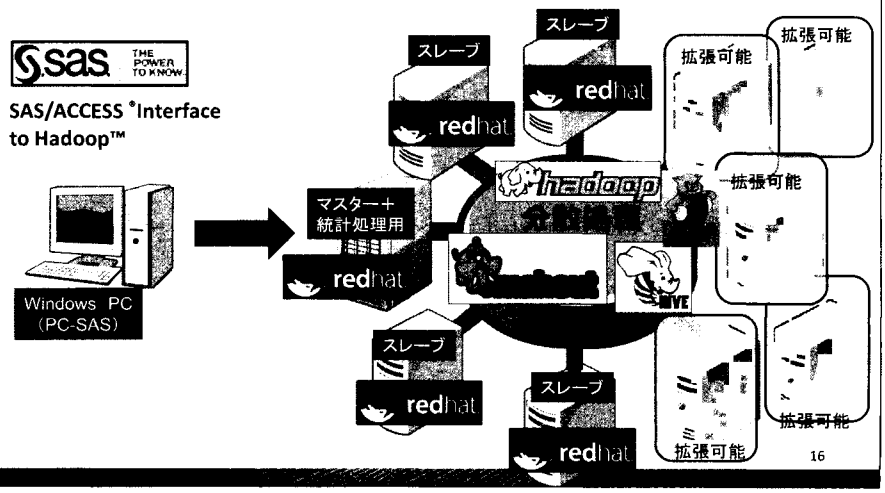
Pig

- PigLatinという言語を使って，HDFS上のデータを操作できる

Mahout

- ビッグデータを用いた機械学習 (レコメンド，クラスタリング，分類) を可能にするライブラリ

塩野義製薬 解析センターの分散処理システム



内容

- 背景
- Hadoop環境の紹介
  - Hadoopとは
  - シオノギのHadoop環境
- SASとHadoopの連携
  - SAS/ACCESS Interface to Hadoopに触れてみて
  - 解析事例

17

SAS/ACCESS Interface to Hadoop

- SASとHadoopをHive経由で接続できる
- 連携のための環境設定がやや複雑
- 連携させられれば使い方はシンプル
  - SQL
  - Hadoop　*proc hadoopからpigでの操作も可能
  - FREQ
  - RANK *PROC RANK in-database processing is not supported by Hadoop.
  - REPORT
  - SORT **The NODUPKEY option is not supported on Hadoop with in-database processing.
  - SUMMARY/MEANS
  - TABULATE
  - etc.

(SAS 9.4 help より)

18

Hadoopに接続して解析を行う(1/2)

Hadoop側で準備しておく

下記コマンドを実行

```
/usr/lib/hive/bin/hive --service hiveserver
```

Point: -hiveconf でHive側の設定を指定可能

ex.) Reducerの数を指定する場合

```
/usr/lib/hive/bin/hive --service hiveserver  -hiveconf mapred.reduce.tasks=25
```


Hadoopに接続して解析を行う(2/2)

SAS側で通常と同様に計算を命令

HDFS上のライブラリを指定

```
option set=SAS_HADOOP_JAR_PATH="D:¥hadoopjar";
libname hd hadoop server = "[XXXXXXXXXX]"
user = [aaaa] password = [xxxxx]   SUBPROTOCOL=hive;
```

計算処理

```
proc means data=hd.simdata mean ;
run;
```

SAS システム

MEANS プロシジャ

変数	ラベル	平均
obsno	obsno	5000000.50
a	a	0.3696910
b	b	0.3098523
c	c	0.3602427
x1	x1	199.4769889
x2	x2	70.99986
x3	x3	6.00008
y	y	0.01057

命令を投げる

処理

Hive

MapReduce

HDFS

結果を受け取る

解析事例(OSIM2)

- OMOP*が公開している，MarketScan® Research Databasesなどの商用データベースをもとにシミュレーションから作成されたデータベース
- Real World Dataの解析手法研究などを目的としている

単純な頻度集計を行って処理速度を比較してみる

- 使用するデータ
  - OSIM2に含まれる薬剤情報のデータ
  - 必要な変数 だけに絞って約30GB程度にした
  - 118,541,933オブザベーション

* OMOP: Observational Medical Outcomes Partnership

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

```
proc freq data=hd.kore_temp2 order=freq;
table CONCEPT_NAME;
run;
```

結果抜粋

SAS システム

FREQ プロシジャ

concept_name

concept_name	度数	パーセント	累積度数	累積パーセント
Acetaminophen	3730371	2.15	3730371	3.15
Amoxicillin	2908652	2.45	6639223	5.60
Hydrocodone	2604931	2.20	9244154	7.80
Azithromycin	2202637	1.86	11446791	9.66
Albuterol	1500506	1.27	12947297	10.92
fluticasone	1249047	1.05	14196344	11.98
Ciprofloxacin	1232418	1.04	15428762	13.02
Clavulanate	1218199	1.03	16646961	14.04
Ibuprofen	1171494	0.99	17818455	15.03
Pseudoephedrine	1165612	0.98	18984067	16.01
Promethazine	1140723	0.96	20124790	16.98
Hydrochlorothiazide	1138580	0.96	21263370	17.94

- Hadoopを使用　:48秒
- 通常のSAS　:4分31秒

本事例では，Hadoopを使うことで一定のメリットが得られた

まとめ

- 世間の流れと同様，医薬関連データの量も増加の一途を辿っており，並列演算処理できる環境が必要となってきている
- Javaプログラマに頼らず，SAS/ACCESS Interface to HadoopをつかってSASプログラマフレンドリーな環境を整えることは選択肢の一つである

今後

- 将来的には，SASプログラマがHadoop環境を意識せずに解析を行えることが理想である
- Hadoopの得意な処理を把握し，SASの処理と使い分けることが重要である

参考文献, Website

- はじめてのHadoop ~分散データ処理の基本から実践まで, 田澤 孝之, 横井 浩, 松井 一比良,技術評論社 (2012).
- Observational Medical Outcomes Partnership, http://omop.org/.

システム管理負荷を軽減させる、SAS BI 運用に関する検討

独立行政法人　国立がん研究センター
青柳吉博

要旨

- SAS BI ServerおよびVDIを用いた利便性・拡張性に優れた業務環境をご紹介します。VDIを中心として業務環境を仮想化し統合されたシステム上で管理することで、システム担当者の負担を最小限にしつつ、耐障害性・セキュリティに優れた業務環境を幅広く運用することが可能です。

- キーワード：医師主導治験、仮想環境、セキュリティ対策、障害対策

内容

- 国立がん研究センターにおける医師主導治験実施体制について
- システム管理面から見たデータマネジメント体制の問題点
- SAS BI Serverを用いたデータマネジメント体制
- システム担当者の負担を軽減させるインフラ環境の構築
- まとめ

国立がん研究センターにおける医師主導治験実施体制について

- トライアルマネジメント　11名
  - 築地　4名
  - 柏　7名
- データマネジメント　12名
  - 築地　7名
  - 柏　5名
- 生物統計　1名（柏のみ）
- システムメンテナンス　1名（柏のみ）
- 監査　2名
  - 築地　1名
  - 柏　1名

上記以外にも、CRCや薬事専門家など100名以上が業務に関与しています。

国立がん研究センターにおける医師主導治験実施体制について

- 医師主導治験実施試験数　21件
- 医師主導臨床試験実施数　34件
- その他（EDC提供など）　3件

国立がん研究センターにおけるデータマネジメント環境（導入前）

システム管理面から見たデータマネジメント体制の問題点

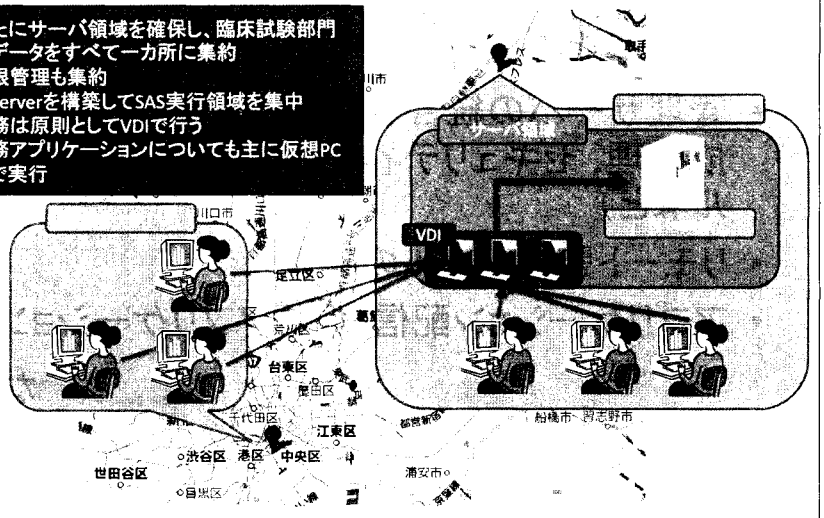
- 築地・柏分かれてサーバが構築されており、相互のデータ参照性は持っていなかったため、事実上協業が不可能だった。
- 業務データの保管場所に関する権限管理が統一されていなかったため、設定が煩雑になっていた。
- データマネジメントはSASで行なっていたものの、実行環境のPCはユーザ自身で管理していたため、PCの障害時や更新時に適切な対応が行えなかった。

SAS BI Serverを用いたデータマネジメント体制

- 新たなサーバ領域を確保して、臨床試験部門の業務データ、個人データ全てをサーバに保管する。(築地・柏ともに一カ所に保管される)
- 上記領域に保管されたデータは統一された権限管理の下で運用される。
- SAS BI Serverを構築しSAS 実行領域は原則BI Server に統合する。
- 接続元PCの環境に依存しないよう作業PCを全てVDI(仮想PC)化する。
- 業務で利用するアプリケーションも原則仮想環境のみで実行する。
- 柏キャンパスにシステム管理者を配置し、システムメンテナンス、権限管理等を集中化する。

SAS BI Server導入後のデータマネジメント体制

- 新たにサーバ領域を確保し、臨床試験部門のデータをすべて一カ所に集約
- 権限管理も集約
- BI Serverを構築してSAS実行領域を集中
- 業務は原則としてVDIで行う
- 業務アプリケーションについても主に仮想PC上で実行



SAS BI Server導入後にシステム管理者が管理するサーバ群

- SAS BI Server
- シェアポイントサーバ
- VDI
- 認証基盤
- ファイルサーバ
- プリントサーバ
- バックアップサーバ
- 監視用サーバ
- 構成管理用サーバ

など計20以上を1人で管理→運用負荷を軽減させるために自動化・省力化が必須

システム担当者の負担を軽減させるインフラ環境の構築

- 認証基盤の統一
- 管理サーバの統合（権限管理、死活監視、構成管理、セキュリティ対策、ヘルプデスクツールなど）
- リモート監視ツールの導入
- アプリケーション配信（仮想アプリケーション）

など

まとめ

- 国立がん研究センターにおける医師主導治験実施体制について説明しました。
- SAS BI ServerとVDIを導入することで、統一された環境で簡便にSASの運用を行う事ができます。
- さらに、管理サーバや認証基盤を統合させるなど、サーバ管理者の負担を最小限にする工夫を行う事で、少人数でのシステム運用が可能となります。

SASを教える・SASを始める
新しいSASの操作方法
SAS® Enterprise Guide のご案内

古賀信二
SAS Institute Japan株式会社

To teach, to start with the new way of using SAS by **SAS® Enterprise Guide**

SHINJI KOGA
SAS Institute Japan Ltd.

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨： 【SASを教える】【SASを始める】方に新しいSASの操作方法をご案内します。

SASは豊富なプロシジャを提供し、SAS言語をマスターすることで、
望む統計手法を自在に実行できます。
しかし、SASが初めての人にとっては、SAS言語を学ぶことも、習得することも大変です。
また、SASユーザにとっても、人材育成・確保などの課題が発生しています。

キーワード：

SAS® Enterprise Guide・SAS Office Analytics

2

SAS(R) 9.3 Procedures by Base SAS & SAS/STAT

	様々な相関係数の計算	CORR
基本統計	度数表・集計表の作成、および集計データに対する解析	FREQ
	基本統計量の算出	MEANS 、SUMMARY
	1変量に対する詳細な分析	UNIVARIATE
データの順位付けと標準化	順位付け	RANK
	標準化	STANDARD
ユーザー定義関数の作成	ユーザー定義関数の作成	FCMP (SAS9)
	条件の悪いデータに対する線形回帰	ORTHOREG
回帰分析	非線形回帰	NLIN
多変量解析	因子分析	FACTOR

この他にも数多くのプロシジャがあります

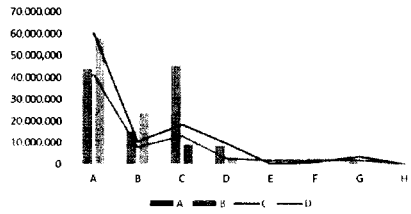
医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

```
/* ---------------------------------------------------------------
  ソートデータセット C:¥SAS¥Data¥SASData¥DEMOSFJ¥sales.sas7bdat
  --------------------------------------------------------------- */

PROC SQL;
            CREATE VIEW WORK.SORTTempTableSorted AS
                        SELECT T."購入金額"n
            FROM ECLIB000.sales as T
;
QUIT;
TITLE;
TITLE1 "分布: 購入金額";
FOOTNOTE;
            ODS EXCLUDE EXTREMEOBS MODES MOMENTS
QUANTILES;

            GOPTIONS htext=1 cells;
            SYMBOL v=SQUARE c=BLUE h=1 cells;
            PATTERN v=SOLID
            ;
PROC UNIVARIATE DATA = WORK.SORTTempTableSorted
                        CIBASIC(TYPE=TWOSIDED
ALPHA=0.05)
                        MU0=0
;
            VAR "購入金額"n;
            HISTOGRAM   "購入金額"n / NORMAL        (
            W=1         L=1             COLOR=YELLOW
MU=EST SIGMA=EST)

                        CFRAME=BLACK CAXES=RED WAXIS=1
CBARLINE=BLACK CFILL=BLUE PFILL=SOLID ;
            ;
/* ---------------------------------------------------------------
  タスクコードの終わりです。
  --------------------------------------------------------------- */
RUN; QUIT;
```



【プログラムがわからない方】

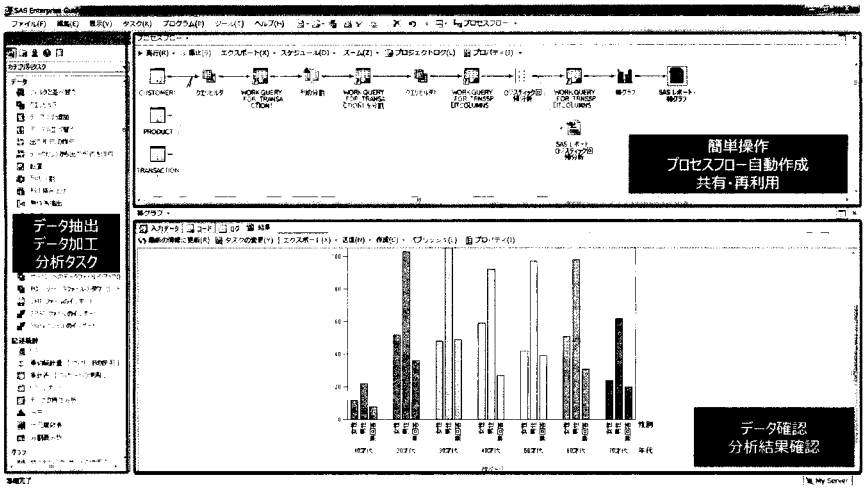
- ✓ 過程がわからない
- ✓ 覚える意欲低下
- ✓ 身近に教わる環境

【プログラムを教える方】

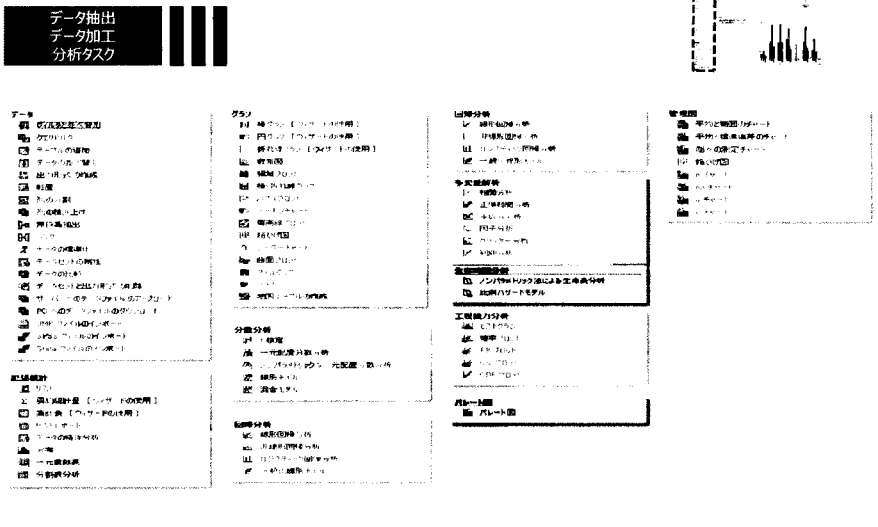
- ✓ 説明が難しい
- ✓ 業務負担の増大
- ✓ 属人化問題

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

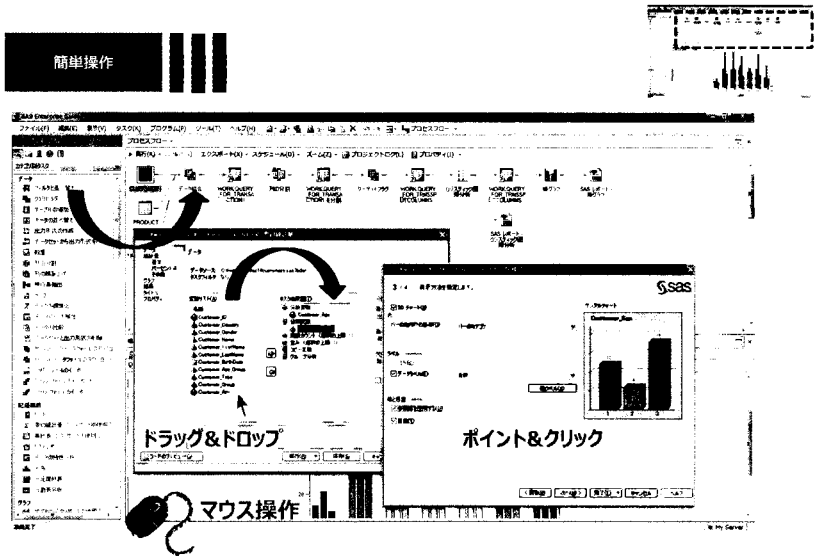
SAS® Enterprise Guide



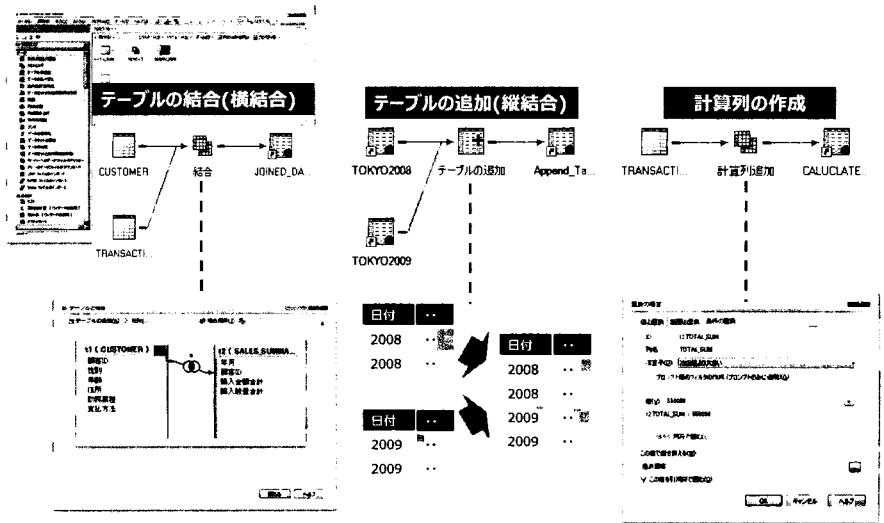
医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014



医療、政府・自治体、大学によるエコシステムの実証　SASユーザー総会　2014

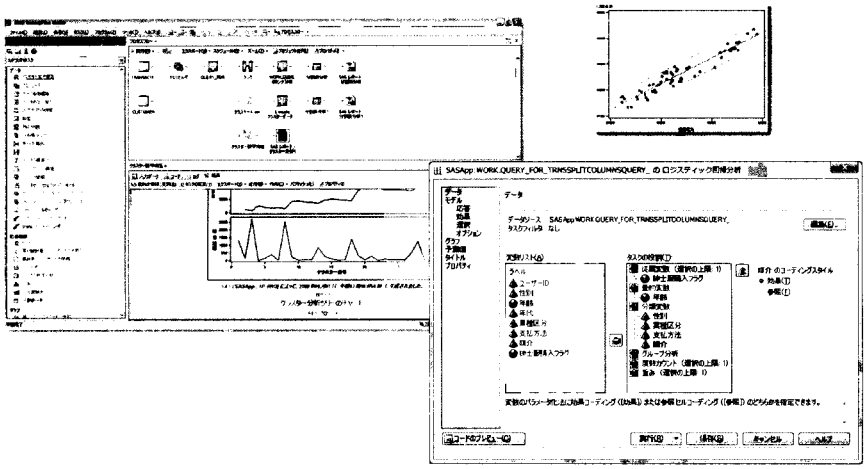


医療、政府・自治体、大学によるエコシステムの実証　SASユーザー総会　2014



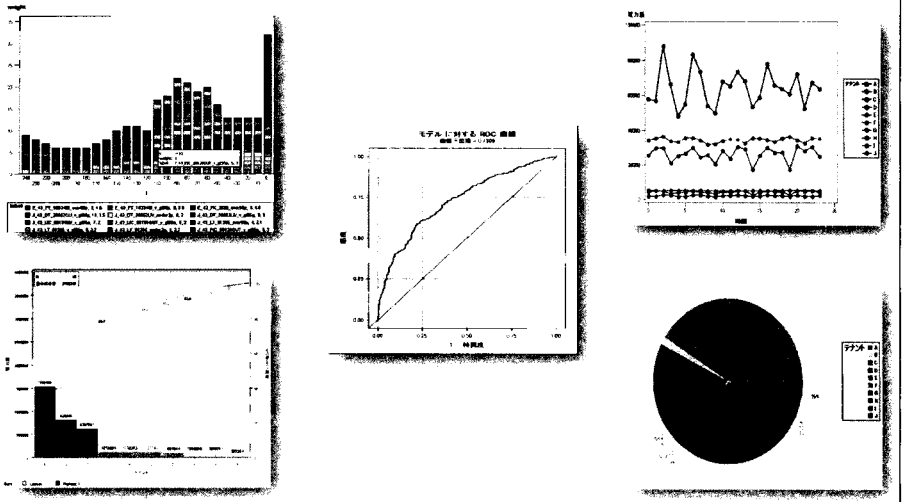
様々なデータ加工メニューを提供

医療、政府・自治体、大学によるエコシステムの実証　SASユーザー総会　2014



集計から回帰分析・多変量解析など統計解析を、簡単なマウス操作で実行

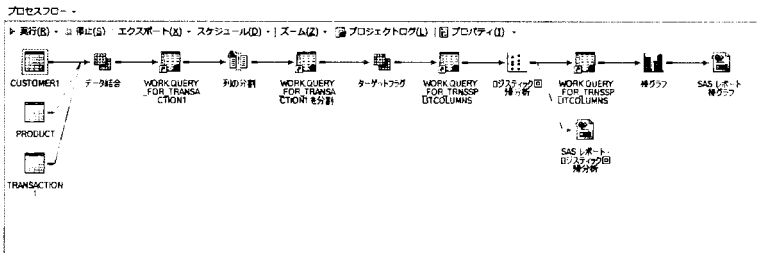
医療、政府・自治体、大学によるエコシステムの実証　SASユーザー総会　2014



グラフ・チャート・表作成の簡単なマウス操作

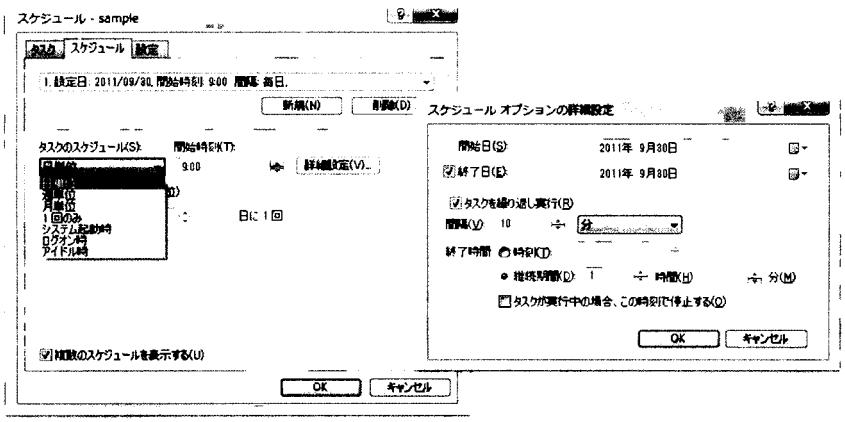
医療、政府・自治体、大学によるエコシステムの実証　SASユーザー総会　2014

プロセスフロー自動生成
共有・再利用



1. 直観的なマウス操作で実行したプロセスをフロー図として自動生成
2. プロセスの編集・条件変更によるプロセス分岐へ対応
3. 再利用・共有へ対応

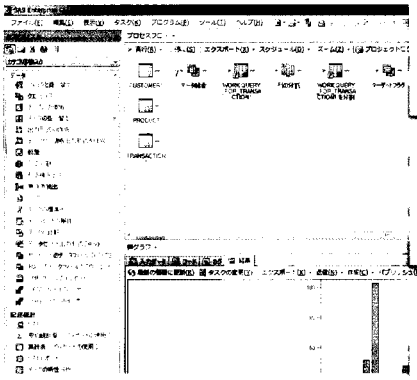
医療、政府・自治体、大学によるエコシステムの実証　SASユーザー総会　2014

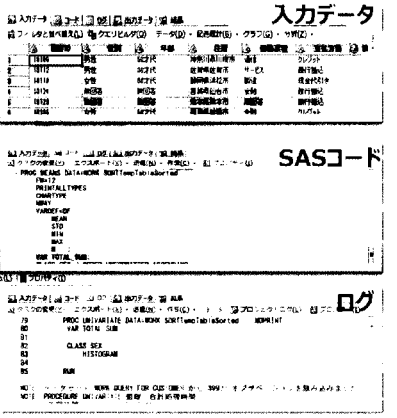


プロセスフローをスケジュール設定でバッチ実行が可能

医療、政府・自治体、大学によるエコシステムの実証　SASユーザー総会　2014

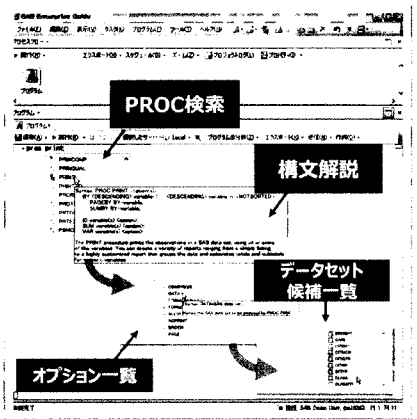
データ確認

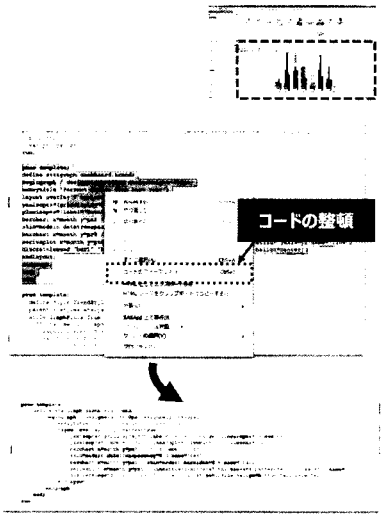




医療、政府・自治体、大学によるエコシステムの実証　SASユーザー総会　2014

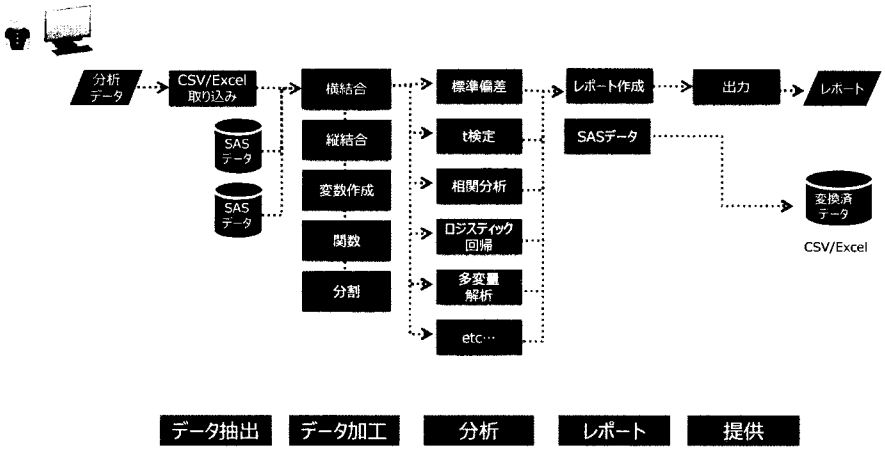
SAS言語と連携





SASプロシジャおよび解説、データセット名の候補情報が自動的に表示されコーディングを支援する

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014



医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

- デモ内容 ： SAS Enterprise Guide基本機能
  データ結合・要約機能・列の分割・変数作成・ロジスティック回帰・プロンプト・エクスポート
- 利用データ
  - 「受注データ」「顧客マスタ」「製品マスタ」

「受注データ」（データ名：TRANSACTION、外部キー：「商品小分類コード」、「ユーザーID」）

	年度	オーダーID	購入金額	単価	購入数量	商品小分類コード	ユーザーID
1	2009	1	990	495	2	81219651	10100
2	2009	1	555	555	1	81079361	10100
3	2009	2	990	495	2	81219549	10100
4	2009	3	375	375	1	81519975	10100
5	2009	4	495	495	1	81219641	10100

「顧客マスタ」（データ名：CUSTOMER、主キー：「ユーザーID」）

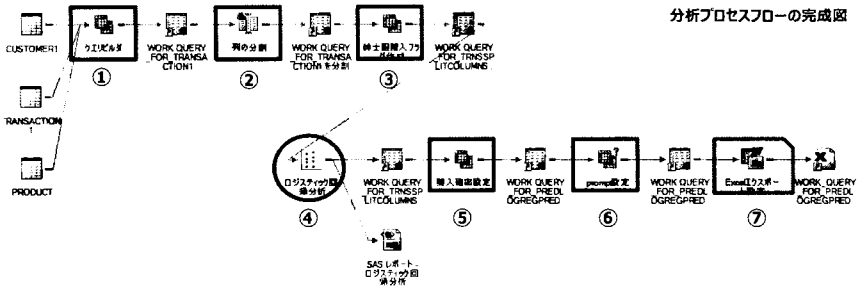
	ユーザーID	性別	年齢	年代	業種区分	支払方法	紹介	住所	登録日
1	10100	男性	37	30才代	通信	クレジット	友人	神奈川県川崎市	2010/10/16
2	10112	男性	54	50才代	サービス	銀行振込	Web	佐賀県佐賀市	2010/07/29
3	10118	女性	19	10才代	製造	現金代引き	友人	静岡県浜松市	2008/05/09
4	10124	無回答	65	60才代	金融	銀行振込	雑誌	宮城県仙台市	2010/09/25
5	10128	女性	69	60才代	製造	銀行振込	Web	熊本県熊本市	2009/07/07

「商品マスタ」（データ名：PRODUCT、主キー：「商品小分類コード」）

	分類ID	商品大分類	商品中分類	商品小分類コード
1	1	Footwear	デイハイキング	81239665
2	1	Footwear	デイハイキング	81239667
3	1	Footwear	デイハイキング	81239669
4	1	Footwear	デイハイキング	81239671
5	1	Footwear	デイハイキング	81239673

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

分析プロセスフローの完成図

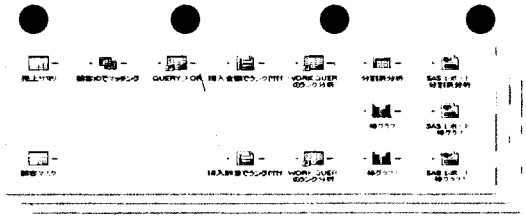


① 「受注データ」「顧客マスタ」「製品マスタ」を結合 （データ加工）
② 縦持ちの商品中分類データを横持ちへ転置 （データ加工）
③ ２値の値を持つ、紳士服購入フラグを作成 （データ加工）
④ 回帰分析で、商品購入に影響する属性を特定し、顧客毎の購入確率を算出 （分析）
⑤ 紳士服購入確率をパーセントへ変更 （データ加工）
⑥ プロンプトの設定で、紳士服購入確率と性別の条件設定 （データ加工）
⑦ ターゲットリストをExcelへエクスポート

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

(Office Analytics)
結果の共有

データ選択　分析用に加工　探索＆分析　結果

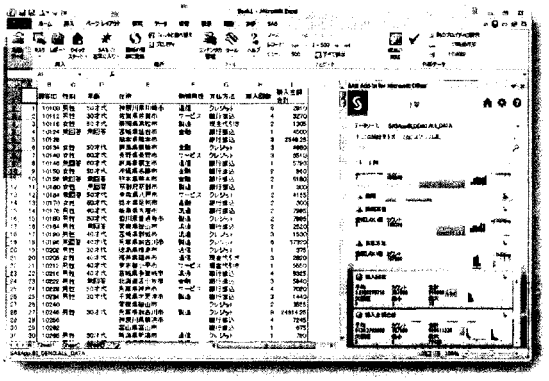


Stored Processとして配布

SAS Enterprise Guideで生成されたプロセスフローをパッキングし、
オフィスソフトからセルフサービス型の配信と共有が可能

※SAS Office Analyticsライセンスが必要

SAS® Office Analytics



1. Microsoft Officeからあらゆるデータへの直接アクセス、分析が可能
2. カスタム処理(*ストアドプロセス)の実行により複雑な処理を実現可能
3. Officeの機能と連携により、レポーティング業務を自動化・効率化

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

『The Power to Know®』

経済分析

基礎自治体のSNSを活用した情報発信の有効性の評価

有馬　昌宏
兵庫県立大学　応用情報科学研究科

Evaluation of Effectiveness of Information Provision by Local Governments through Social Networking Services

Masahiro Arima
Graduate School of Applied Informatics, University of Hyogo

要旨

近年，Twitter や Facebook などのソーシャルメディアの普及を受け，また東日本大震災で果たしたソーシャルメディアの機能の有効性を受け，自治体においても，①情報の新たな発信手段の確保，②双方向コミュニケーションの特性を活かしての行政の説明責任や広聴による住民の地域行政へのより深い関与，③地域キャラクター（ゆるキャラ）からの情報発信を利用しての観光・訪問客の増加，などを目的に，広報・公聴の媒体としてソーシャルメディアを導入する動きが見えはじめてきている．しかし，自治体でのソーシャルメディアによる広報公聴は，より多くの住民へのリーチやより多くの住民によるエンゲージメント，さらには地域理解やコミュニケーションによる QOL（Quality of Life）の向上に資する可能性がある一方で，情報漏洩・プライバシー侵害・信頼性損失・風評被害などのリスクも含んでおり，実際に住民がどれだけの効用を得られるかについて，さらにはリスクの評価についての定量的分析は全く行われてはいない．そこで，自治体でのソーシャルメディアを活用した広報・公聴が住民の効用水準をどれだけ増大させうるか定量的に把握するために，選択実験と仮想市場法（CVM：Contingent Valuation Method）の手法を用い，2014 年３月から４月にかけて全国を対象とするウェブ調査を実施して，5,005 サンプルからの回答を得た．本発表では，この回答データに基づき，自治体からのソーシャルメディアを活用した広報・公聴活動が住民にどのように受容され，どれだけの効果をもたらすのかについて，JMP（Ver.11.0.0）を利用しての分析と定量評価を試みる．

キーワード：基礎自治体，情報発信，SNS，有効性評価，ウェブ調査，選択実験，仮想市場評価法

１．はじめに

近年，Twitter や Facebook などのソーシャルメディアの普及を受け，また東日本大震災で果たしたソーシャルメディアの機能の有効性を受け，自治体においても，①情報の新たな発信手段の確保，②双方向コミュニケーションの特性を活かしての行政の説明責任や広聴による住民の地域行政へのより深い関与，③地域キャラクター（ゆるキャラ）からの情報発信を利用しての観光・訪問客の増加，などを目的に，広報・公聴の媒

体としてソーシャルメディアを導入する動きが見えはじめてきている．実際，佐賀県武雄市では，2011 年８月１日より公式ウェブサイトの全てを Facebook に移行しており，災害時の情報発信のチャネル確保や地域への関心の喚起などの目的で武雄市に追随する自治体も動向を見極めて今後は増加していくものと思われる．

しかし，自治体でのソーシャルメディアによる広報公聴は，より多くの住民へのリーチやより多くの住民によるエンゲージメント，さらには地域理解やコミュニケーションによる QOL（Quality of Life）の向上に資する可能性がある一方で，情報漏洩・プライバシー侵害・信頼性損失・風評被害などのリスクも含んでいる．また，これまでにも，地域 SNS（Social Networking Service）への期待と「立ち枯れ」化という現状があり，これらのソーシャルメディアを活用した地域の振興や自治の発展への期待が，地域 SNS と同じような道を辿ることは避けなければならない．そのためには，ソーシャルメディアの自治体による活用の可能性（自治体内の地域住民と自治体外の潜在訪問者にもたらすメリット）と課題（プライバシー侵害などのデメリット）を定量的に把握しておき，コストと効果の観点からの事前評価をきちんと行っておく必要がある．

ところが，このような自治体によるソーシャルメディアの利活用に伴うメリットとデメリットの定量評価に関する研究は行われてはいない．そこで，本研究では，自治体でのソーシャルメディアを活用した広報・公聴が住民の効用水準をどれだけ増大させうるかを，選択実験と仮想市場法（CVM：Contingent Valuation Method）の手法を用い，全国を対象とするウェブ調査を実施して，定量的に評価することを試みる．

本研究により，自治体からのソーシャルメディアを活用した広報・公聴が，自治体内の住民と自治体外からの潜在来訪者に分けて，どれだけの効果をもたらすかの定量評価が初めて可能となる．その結果，ウェブ調査というサンプルの偏りの面での問題はあるものの，逆にウェブ調査であることを活かして，当面のソーシャルメディアの潜在的積極利活用者であるインターネット利用者を調査対象とする調査結果から自治体のソーシャルメディアによる広報・公聴活動のメリットとデメリットを金銭的に定量評価することが可能となり，ソーシャルメディアを利活用する施策立案および政策評価に向けての基礎的データを提供できるものと考えられる．

２．研究調査の方法とその概要

自治体におけるソーシャルメディアの一つである地域 SNS（Social Networking Service）の利活用は，2004 年の熊本県八代市の「ごろっとやっちろ」の成功事例を受け，さらには総務省や財団法人地方自治情報センターなどからの支援もあり，2005 年以降は全国の自治体に広がる傾向にあったが，2010 年頃には，運営コストを上回る効果を生んでいるかという経済性の問題やユーザ数やコミュニケーション数の減少による「立ち枯れ化」の問題が顕在化しはじめていた（庄司（2009））．このような状況の中，Twitter や Facebook などの民間が運営する新しいタイプの SNS が提供されて普及しはじめ，これらのソーシャルメディアの民間企業や NPO 法人における利活用の成功事例，特に東日本大震災での SNS の果たした役割とその有効性の認識に基づき，自治体が自前で地域 SNS や公式ウェブサイトを提供するのではなく，Twitter や Facebook などのソーシャルメディアをプラットフォームとして，より多くの住民へのリーチとより多くの住民からのエンゲージメントを目標に，積極的に活用されていくであろうことが予想されている．

一般に，自治体でのソーシャルメディアの利活用は，谷本（2013）によれば，①ブロードキャスト型，②キャラクター型，③限定コミュニティ型に分類されるが，本研究では，まず，全国 1,742 の市区町村ならびに 47 都道府県の公式ウェブサイトを対象とした目視での悉皆調査を行うことで，自治体のソーシャルメディア利用の有無と利活用のされ方の分類による現状把握を行った．その結果，2013 年 12 月８日時点で，全国

の 1,719 の市町村では，14.5%の 250 市町村でツイッターによる情報発信が行われており，21.2%の 365 市町村ではフェイスブックによる情報発信が行われていることが確認できた．なお，我々の目視調査では，自治体からフェイスブックとツイッターで発信される情報が防災・安心・安全に関連する情報に特化したものであるかどうかもチェックしているが，防災・安心・安全に関する情報発信に特化してツイッターから発信しているのは 18 自治体，フェイスブックから発信しているのは 14 自治体であった．

一方で，ソーシャルメディアを介しての自治体による情報発信や情報交流について，住民がどのように利用し，どのように評価しているかを把握するために，ウェブ調査を実施した．ウェブ調査では，回答者に偏りが生じることは十分に承知しているが，自治体のソーシャルメディア活用による情報発信ならびに双方向コミュニケーションの最初のターゲットと想定されるのは，インターネットやソーシャルメディアの現時点での積極的利用者である．したがって，ウェブ調査であるがゆえに，自治体によるソーシャルメディアでの情報発信の受容の実態の把握や自治体のソーシャルメディアの利活用におけるメリットならびにデメリットの住民による評価が適切に行われると期待できるのではないかと考えたのである．

ウェブ調査の実施にあたっては，筆者も学識経験者として参画していた地方自治情報センター（2013）などの類似の調査の調査票を参考にしながら，①住所（都道府県・市区町村・郵便番号）・性別・年齢，②居住年数，③住居の形態と所有関係，④世帯人員，⑤家族構成，⑥職業，⑦インターネットの接続状況，⑧利用している機器，⑨SNS の知名と SNS の過去と現在の利用状況・利用頻度・利用 SNS および利用しない理由，⑩インターネットの利用年数，⑪インターネットでの商品・サービスの購買状況，⑫市区町村からの広報紙の閲覧頻度，⑬市区町村からのチラシや回覧板での知らせの有無と閲覧頻度，⑭居住している市区町村のウェブサイトの閲覧頻度，⑮居住している市区町村のソーシャルメディアでの情報発信の状況と閲覧状況，⑯居住していない市区町村のソーシャルメディアでの情報発信の閲覧経験，⑰20 分野に分類した地域の情報項目ごとの情報獲得状況と 11 のメディア別の情報獲得媒体，⑱情報の提供媒体と提供内容と利用料の３属性の各水準の組み合わせによる仮想の地区町村からの情報提供の仕組みに対する利用するかしないかの２項選択と５段階評価と順位づけによる評価，⑲自治体がソーシャルメディアを利用することで発生すると思われる情報漏洩や無断書き込みの恐れのある情報および名誉棄損に対して求める損害賠償額と市区町村が賠償支払いのために支払ってもよいと考える保険料の額，⑳自治体のソーシャルメディアによる情報発信への賛否，の 20 の設問から構成される調査票を設計した．その上で，サンプルについては，総サンプル数は 5,000 以上，男女でそれぞれ 2,200 サンプル以上，年齢は 20 歳以上を対象として，30 歳未満，30 歳代，40 歳代，50 歳代，60 歳以上でそれぞれ 450 サンプル以上，47 の各都道府県のうち山梨，佐賀，福井，徳島の各県は 20 サンプル以上，高知，島根，鳥取の各県は 15 サンプル以上，それ以外の都道府県は 30 サンプル以上を確保するという条件で株式会社データーサービスに調査を委託し，「自治体からのソーシャルメディアによる情報発信に関するウェブ調査」というタイトルで 2014 年３月 15 日から応募型で調査を開始し，サンプルの条件が満たされた４月 25 日に調査を終了した．

ウェブ調査で得られた有効サンプルの数は 5,005 であり，都道府県別では，613 サンプルの東京都，413 サンプルの大阪府，402 サンプルの神奈川県がサンプル数の多い都府県であり，22 サンプルの佐賀県と鳥取県，23 サンプルの福井県がサンプル数の少ない県となっている．サンプルの性別と年齢と職業ならびにインターネットの経験年数別の構成比率は図 1 に示すとおりで，男女の比率はほぼ等しいが，年齢別に見ると 50 歳以上は男性が多く，40 歳未満は女性が多いというサンプルの偏りが見られる．また，インターネットの経験年数では，「10 年以上」が 73.6%で最も多く，男女の比率もほぼ等しいが，経験年数が 10 年未満では女性の比率が高いという傾向が見られる．

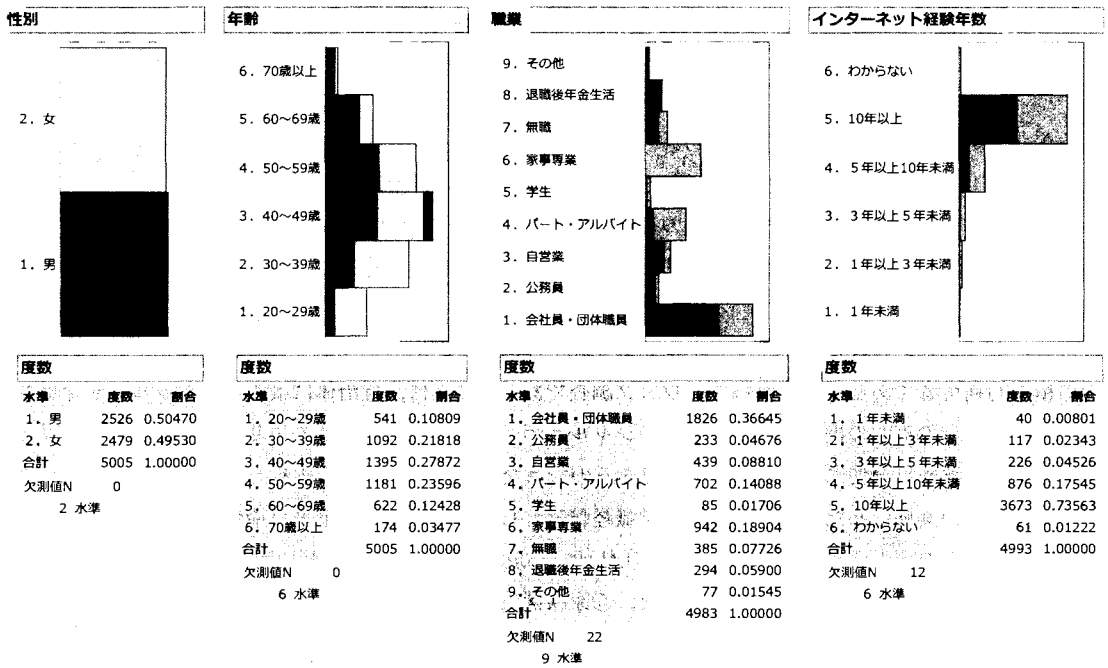


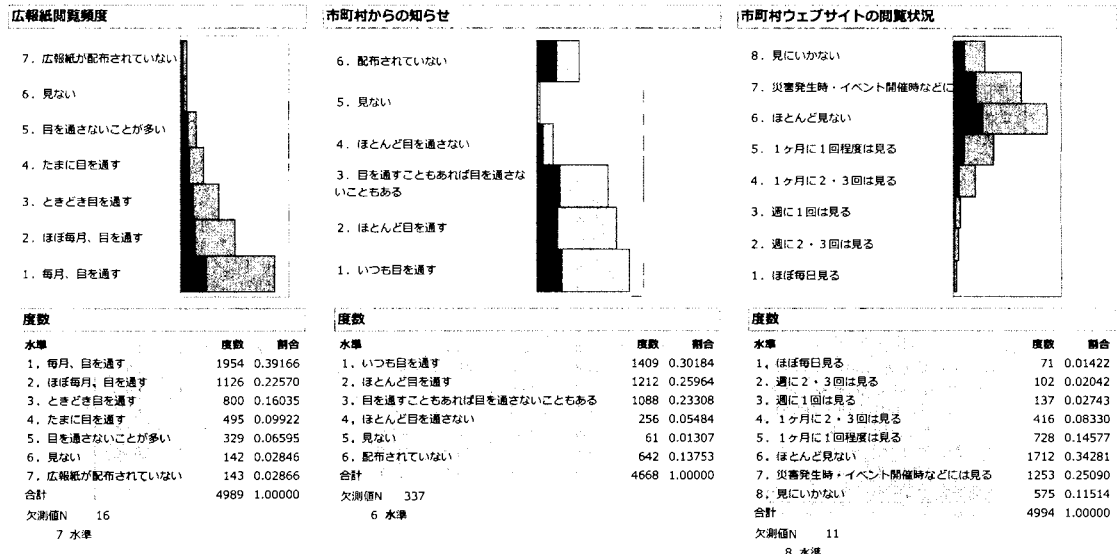
図１　自治体からのソーシャルメディアによる情報発信に関するウェブ調査のサンプル構造

なお，回答者が利用している機器は，PC が 94.0%，タブレット端末が 17.6%，スマートフォンが 41.8%，携帯電話が 41.2%（複数回答なので合計は 100%を超える）となっている．

３．自治体から発信される情報の入手の現状

自治体からは，住民に向けて，さまざまな媒体を通じて，様々な情報が提供されている．本研究では，自治体が住民向けの情報を発信する媒体として，①広報紙，②議会だより，③チラシ・新聞折り込み，④回覧板，⑤ホームページ，⑥ツイッター，⑦フェイスブック，⑧防災行政無線，⑨テレビ・ラジオ，⑩地域のケーブルテレビ，⑪新聞，の 11 の媒体に分類し，自治体から提供されている情報を，①市区町村の総合計画・施策情報，②財政情報（予算や決算など），③議会情報（開催日・議決事項など），④イベント情報，⑤保険情報（健康保険や介護保険），⑥福祉情報（高齢者や児童の福祉の制度など），⑦戸籍・住民票情報（申請・届出など），⑧統計情報（地区別人口など），⑨健康情報（インフルエンザ発生状況など），⑩安心・安全情報（ひったくりや痴漢など），⑪平時の防災情報（消防・救急を含む），⑫災害時の防災情報（警報や避難情報など），⑬公共工事入札情報，⑭教育情報（幼稚園・小中学校関係の情報），⑮文化・歴史情報（文化財や郷土資料など），⑯ごみ収集情報（回収日や回収場所），⑰公共施設情報（開館日や場所），⑱観光情報（観光スポットの紹介など），⑲公共工事実施情報，⑳地域情報（新聞・テレビ報道も含む地域の話題），の 20 の分野に分類して，情報の入手状況を設問している．

自治体においては，広報・広聴活動の主たる媒体は広報紙（議会に関しては議会だより）やチラシ（回覧板，新聞の折り込みチラシ，ポストへ投函される案内チラシなど）であり，インターネットの普及に伴って公式のウェブサイトを設けてホームページからの情報発信も行われている．そして，近年のソーシャルメディアの利用の広がりに合わせて，ツイッターやフェイスブックでの情報発信を行う自治体も増え始めている



度数

水準	度数	割合
1，毎月、目を通す	1954	0.39166
2，ほぼ毎月、目を通す	1126	0.22570
3，ときどき目を通す	800	0.16035
4，たまに目を通す	495	0.09922
5，目を通さないことが多い	329	0.06595
6，見ない	142	0.02846
7，広報紙が配布されていない	143	0.02866
合計	4989	1.00000

欠測値N 16
7 水準

度数

水準	度数	割合
1，いつも目を通す	1409	0.30184
2，ほとんど目を通す	1212	0.25964
3，目を通すこともあれば目を通さないこともある	1088	0.23308
4，ほとんど目を通さない	256	0.05484
5，見ない	61	0.01307
6，配布されていない	642	0.13753
合計	4668	1.00000

欠測値N 337
6 水準

度数

水準	度数	割合
1，ほぼ毎日見る	71	0.01422
2，週に2・3回は見る	102	0.02042
3，週に1回は見る	137	0.02743
4，1ヶ月に2・3回は見る	416	0.08330
5，1ヶ月に1回程度は見る	728	0.14577
6，ほとんど見ない	1712	0.34281
7，災害発生時・イベント開催時などには見る	1253	0.25090
8，見にいかない	575	0.11514
合計	4994	1.00000

欠測値N 11
8 水準

図２　自治体からの情報の閲覧状況（広報紙・チラシ・ホームページ）

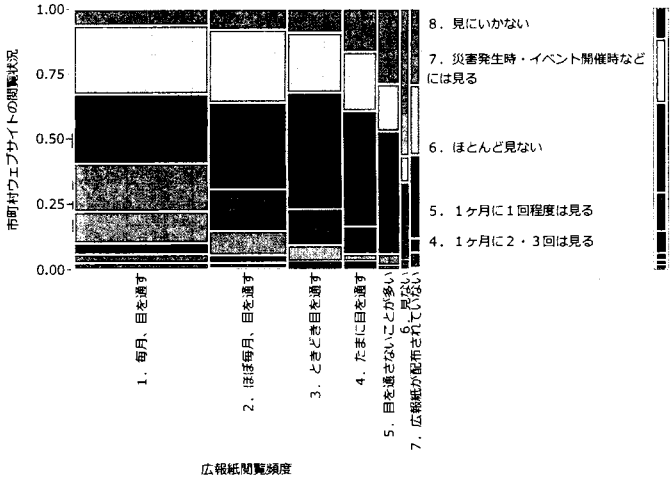
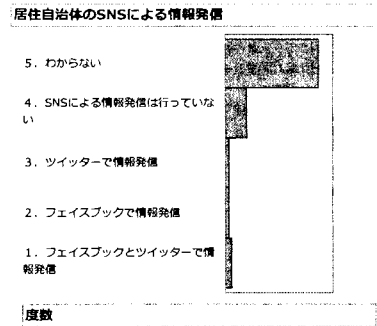


図３　広報紙の閲覧頻度とホームページの閲覧状況との関係



度数

水準	度数	割合
1．フェイスブックとツイッターで情報発信	265	0.05295
2．フェイスブックで情報発信	133	0.02657
3．ツイッターで情報発信	146	0.02917
4．SNSによる情報発信は行っていない	837	0.16723
5．わからない	3624	0.72408
合計	5005	1.00000

欠測値N 0
5 水準

図４　ＳＭでの情報提供の認識

のが現状である．これらの５つの媒体（広報紙，チラシ，ホームページ，ツイッター，フェイスブック）のうち，広報紙とチラシとホームページを通じての情報入手の有無や頻度に関しての回答結果をまとめたのが図２であり，自治体から住民へ届けられるという意味でプッシュ型の紙媒体である広報紙において「毎月，目を通す」と「ほぼ毎月，目を通す」を併せると 61.7%，同じくプッシュ型の紙媒体であるチラシを通す」と「ほとんど目を通す」を併せると 56.1%であるのに対して，住民がアクセスをしなければならないという意味でプル型のホームページは「ほぼ毎日見る」から「１ヶ月に１回程度は見る」までの比率を併せても 29.1%であり，広報紙の閲覧頻度とホームページの閲覧状況との間のモザイク図である図３に示すように，広報紙をよく見る回答者がホームページもよく閲覧するという補完的な関係が窺える結果となっており，ホームページが紙媒体を代替するには至っていないというのが現状である．このような状況に対して，ソーシャルメディア（ＳＭ）を通じての情報入手に関しては，図４に示すように，居住している自治体がフェイスブックまたはツイッターあるいはその両方によって情報発信をしていると認識している回答者は 10.9%であり，自

表１　20の地域情報の分野別の情報入手状況（上は回答者の属性別，下はメディア別）

	全体	性別		年齢					
		男性	女性	20歳代	30歳代	40歳代	50歳代	60歳代	70歳以上
1 市区町村の総合計画・施策情報	67.4	69.4	65.4	49.7	60.7	65.8	73.5	80.7	89.1
2 財政情報（予算や決算など）	49.1	51.1	47.0	31.8	41.8	46.2	54.0	64.3	83.3
3 議会情報（開催日・議決事項など）	44.6	46.6	42.6	29.8	35.2	40.2	51.1	62.2	78.7
4 イベント情報	77.9	75.7	80.2	69.3	76.6	78.2	78.4	83.1	89.1
5 保険情報（健康保険や介護保険）	45.0	42.4	47.7	36.6	40.4	42.9	44.6	58.4	73.0
6 福祉情報（高齢者や児童の福祉の制度など）	48.5	45.1	52.0	36.8	44.7	46.2	49.2	61.7	74.7
7 戸籍・住民票情報（申請・届出など）	42.5	38.5	46.5	36.8	40.8	41.6	42.3	47.3	60.9
8 統計情報（地区別人口など）	38.1	38.7	37.4	28.7	32.5	35.5	40.8	48.9	66.1
9 健康情報（インフルエンザ発生状況など）	49.0	45.4	52.6	40.3	45.0	47.0	48.3	60.6	79.9
10 安心・安全情報（ひったくりや痴漢など）	44.8	41.4	48.3	37.2	41.1	43.4	43.7	55.9	71.3
11 平時の防災情報（消防・救急を含む）	42.0	41.6	42.4	32.7	35.3	39.5	45.5	52.6	70.7
12 災害時の防災情報（警報や避難情報など）	48.8	47.7	49.9	39.0	42.5	48.5	51.1	58.8	69.0
13 公共工事入札情報	24.0	23.0	25.1	20.7	22.6	22.4	24.9	27.8	36.8
14 教育情報（幼稚園・小中学校関係の情報）	38.5	34.4	42.6	35.7	39.9	37.8	35.6	41.0	53.4
15 文化・歴史情報（文化財や郷土資料など）	38.9	38.6	39.2	29.6	31.5	35.6	41.7	53.9	69.5
16 ごみ収集情報（回収日や回収場所）	74.2	71.1	77.3	63.6	70.8	75.2	76.4	79.4	86.2
17 公共施設情報（開館日や場所）	51.9	48.0	55.8	40.5	48.9	51.7	53.0	58.8	74.7
18 観光情報（観光スポットの紹介など）	47.4	43.5	51.4	44.9	45.7	46.3	46.0	53.2	63.8
19 公共工事実施情報	27.4	26.0	28.8	25.5	25.7	24.8	27.2	33.4	44.3
20 地域情報（新聞・テレビ報道も含む地域の話題）	52.2	49.4	55.1	47.5	48.6	51.0	52.4	58.7	73.6
サンプル数	5,005	2,526	2,479	541	1,092	1,395	1,181	622	174

	全体	広報誌	議会だより	チラシ折り込み	回覧板	ホームページ	ツイッター	フェイスブック	防災行政無線	テレビラジオ	地域のケーブルテレビ	新聞
1 市区町村の総合計画・施策情報	67.4	52.7	14.6	6.5	11.2	10.8	0.4	0.6	0.9	5.5	2.1	12.0
2 財政情報（予算や決算など）	49.1	36.4	11.2	1.9	3.1	4.3	0.1	0.1	0.1	2.2	0.7	5.2
3 議会情報（開催日・議決事項など）	44.6	36.4	11.2	1.9	3.1	4.3	0.1	0.1	0.1	2.2	0.7	5.2
4 イベント情報	77.9	23.3	20.6	1.9	2.5	3.5	0.2	0.1	0.3	1.9	1.0	3.4
5 保険情報（健康保険や介護保険）	45.0	57.4	3.2	14.3	16.2	16.4	0.8	1.0	1.0	10.0	4.3	13.9
6 福祉情報（高齢者や児童の福祉の制度など）	48.5	32.2	1.8	3.2	4.2	9.2	0.1	0.2	0.2	2.1	0.6	3.9
7 戸籍・住民票情報（申請・届出など）	42.5	35.9	2.3	2.7	5.2	10.8	0.1	0.2	0.2	2.1	0.8	3.9
8 統計情報（地区別人口など）	38.1	28.2	2.0	1.3	1.8	7.4	0.2	0.2	0.1	1.3	0.3	3.0
9 健康情報（インフルエンザ発生状況など）	49.0	29.0	1.5	2.4	5.6	9.7	0.5	0.3	0.6	7.3	1.3	10.4
10 安心・安全情報（ひったくりや痴漢など）	44.8	22.1	1.6	3.5	13.0	5.9	0.6	0.3	2.1	4.8	1.4	7.5
11 平時の防災情報（消防・救急を含む）	42.0	25.3	1.6	2.3	7.2	8.0	0.4	0.2	3.2	4.1	1.2	4.7
12 災害時の防災情報（警報や避難情報など）	48.8	24.8	1.6	2.3	6.5	11.2	1.0	0.4	5.8	8.7	2.0	5.7
13 公共工事入札情報	24.0	15.2	2.0	1.1	1.8	4.1	0.1	0.1	0.1	1.0	0.5	2.7
14 教育情報（幼稚園・小中学校関係の情報）	38.5	26.4	1.8	2.2	5.5	8.7	0.3	0.2	0.2	2.2	1.0	4.6
15 文化・歴史情報（文化財や郷土資料など）	38.9	28.2	1.4	2.3	3.2	7.9	0.1	0.2	0.2	3.3	1.3	6.1
16 ごみ収集情報（回収日や回収場所）	74.2	45.9	2.2	10.4	18.2	16.8	0.2	0.2	0.5	1.1	0.6	2.5
17 公共施設情報（開館日や場所）	51.9	34.7	1.6	2.7	5.5	17.4	0.3	0.3	0.4	1.9	1.0	3.6
18 観光情報（観光スポットの紹介など）	47.4	27.5	1.4	4.4	3.2	14.3	0.5	0.5	0.3	7.5	2.4	9.9
19 公共工事実施情報	27.4	17.3	2.1	1.6	3.6	4.2	0.1	0.2	0.2	1.2	0.6	2.7
20 地域情報（新聞・テレビ報道も含む地域の話題）	52.2	28.1	2.0	4.9	6.4	7.9	0.6	0.6	0.9	15.3	6.0	19.0
サンプル数	5,005	5,005	5,005	5,005	5,005	5,005	5,005	5,005	5,005	5,005	5,005	5,005

治体のフェイスブックに登録して閲覧している回答者は1.3%，登録はしてはいないがフェイスブックを閲覧したことがある回答者は1.0%，自治体のツイッターのフォロワーになっている回答者は0.9%，フォロワーではないがツイッターを閲覧したことがある回答者は1.1%にとどまっており，インターネットやスマートフォンの積極的ユーザと考えられる本ウェブ調査の回答者の間でも，ソーシャルメディアを通じて自治体からの情報を入手している回答者の比率は非常に低いというのが現状となっている．

また，表１には，20の分野に分類した地域情報の分野別の情報取得状況を回答者の属性別（性別と年齢別）と11のメディア別にまとめて示している．表１から，住民がよく入手している情報は，①イベント情報（77.9%），②ゴミ収集情報（回収日や回収場所）（74.2%），③市区町村の総合計画・施策情報（67.4%），④地域情報（新聞・テレビ報道も含む地域の話題）（52.2%），⑤公共施設情報（開館日や場所）（51.9%）の順であり，一番低い公共工事入札情報でも24.0%の回答者が情報を入手していると回答している．

入手する情報に性別と年齢別で差があるかどうかについては，性別に関してはFisherの正確検定を，年齢別に関してはPeasonのχ^2検定を適用して検定を行った．その結果，年齢別では，表１からも年齢が高くなるにつれて情報の入手率が高くなる傾向が窺えるが，20の全ての情報分野で，１%有意水準で有意差が認め

ケース	情報の提供媒体	情報の提供内容	情報の提供費用	(1) 利用の可否		(2) 評価					(3) 順位		
				利用する	利用しない	非常に良い	まあまあ良い	どちらともいえない	あまり良くない	良くない	1位	4位	7位
											2位	5位	8位
											3位	6位	9位
1	ツイッター	動画付き	100円／月	○	○	○	○	○	○	○	選択		
2	フェイスブック	文字情報のみ	200円／月	○	○	○	○	○	○	○	選択		
3	ウェブサイト	静止画付き	無料	○	○	○	○	○	○	○	選択		
4	ツイッター	文字情報のみ	無料	○	○	○	○	○	○	○	選択		
5	フェイスブック	静止画付き	100円／月	○	○	○	○	○	○	○	選択		
6	ウェブサイト	動画付き	200円／月	○	○	○	○	○	○	○	選択		
7	ツイッター	静止画付き	200円／月	○	○	○	○	○	○	○	選択		
8	ウェブサイト	文字情報のみ	100円／月	○	○	○	○	○	○	○	選択		
9	フェイスブック	動画付き	無料	○	○	○	○	○	○	○	選択		

図５　選択実験の９つのプロファイルと回答画面

られている．

性別については，統計情報，平時の防災情報，災害時の防災情報，公共工事入札情報，文化・歴史情報の５分野で有意差は認められなかったが，公共工事実施情報では５％有意水準で，それ以外の情報分野では１％有意水準で男女間の情報入手率に有意差が認められている．なお，男女間で有意差が認められた 15 の情報の分野の中で男性の入手率が女性の入手率を上回っている分野は，市区町村の総合計画・施策情報，財政情報，議会情報の３つの分野だけとなっている．

分野別の情報を入手するかどうかには，性別や年齢だけでなく，家族構成や職業などの属性も影響を及ぼすと考えられるが，本稿では紙数の関係で割愛する．

４．ソーシャルメディアによる情報発信のメリットとデメリットの評価

自治体でのソーシャルメディアを利用しての広報・広聴（公聴）活動は，谷本（2011）が指摘する即時性，公開性，双方向，コスト削減というメリットに加えて，情報漏洩・プライバシー侵害・信頼性損失・風評被害や住民間での情報格差の拡大などのデメリットも存在する．しかし，これらのメリットとデメリットが定量的に評価されたことはなかった．

そこで，本研究では，メリットについては選択実験の手法を適用し，デメリットについては CVM（Contingent Valuation Method，仮想市場評価法）の手法を適用して．メリットとデメリットの定量評価を試みることとした．

具体的には，メリットの評価に関しては，①プラットフォーム（フェイスブック，ツイッター，ウェブページ），②提供情報の内容（テキストのみ，写真やイラストなどの静止画付き，動画付き），③費用分担（無料，月額 100 円，月額 200 円）の３属性（各属性の水準数は３）による選択実験により，全部で 27 の組み合わせの中から実験計画法のカスタム計画を用いて選択した９つのプロファイルを提示し，各プロファイルに対して，「利用する」と「利用しない」の２項選択型の選択肢と「非常に良い」，「まあまあ良い」，「どちらとも言えない」，「あまり良くない」，「良くない」の５段階評価法の選択肢に回答してもらうとともに，９つのプロファイルに対して１位から９位までの順位づけを行ってもらい，これらの回答データを用いて各属性・各水準の部分効用値の推定を行うとともに，金銭評価を行った．なお，提示した９つのプロファイルと質問の形式は，図５に示すとおりである．

効用関数の推定に際しては，土木学会土木計画学研究委員会（1995）を参考に，回答者の固定効果の存在を考慮して，条件付ロジットモデルを適用して推定を行った．推定に際しては，まず，プロファイル j が「利用する」として選択されたときの効用を U_j として，次のように確率効用関数を定式化する．

$$U_j = V_j + e_j = \sum_{k=1}^{K} \beta_j x_{kj} + \sum_{l=1}^{L} \gamma_l z_l + \varepsilon_j \quad (1)$$

ここで，V_j はプロファイル j から得られる効用の観察可能な部分を，x_{kj} はプロファイル j の k 番目の属性のダミー変数，z_l は l 番目の回答者の個人属性のダミー変数，β_j , γ_l はそれぞれの推定パラメータを表す．誤差項 ε_j は効用の観察不可能な部分であり，プロファイルと個人について互いに独立なガンベル分布に従うものと仮定する．このとき，プロファイル j が選択される確率 Pj は，

$$P_j = \exp(V_j) / \Sigma_j \exp(V_j) \quad (2)$$

と表わせる．この確率をパラメータの関数と見なして尤度関数を作り，JMP の名義ロジスティック回帰を適用することにより，最尤法によりパラメータを推定する．なお，プロファイルに関する属性のダミー変数の基準としたのは，プラットフォーム（情報の提供媒体）では「ホームページ（ウェブサイト）」，情報の提供内容では「テキスト（文字情報）のみ」，情報の費用負担（提供費用）では「無料」である．また，回答者の固定効果の候補としてモデルに組み込んだのは性別，年齢，職業の３つの属性であり，性別は「女性」を，年齢は「40 歳代」を，職業は「公務員」を基準としている．

９つのプロファイルへの利用の可否質問および３つの固定効果に関する個人属性に関する質問の全てに回答のあった 4,899 サンプルを用いた係数の推計結果を示したものが表２である．表２より，固定効果を組み込まないモデル１，性別と年齢の固定効果を組み込んだモデル２，職業の固定効果を組み込んだモデル３のいずれにおいても，提供内容の「動画付き」を除く全ての属性の水準が１％有意水準で有意となっている．プラットフォームとしては，ウェブサイトと比較して，ツイッターもフェイスブックも推定された係数はプラスとなっており，利用を阻害する要因として効いていることがわかる．提供内容に関しては，静止画付きの係数はマイナスで，利用を促進する要因として効いている．負担費用に関しては，推定された係数はプラスの大きな値となっており，利用を大きく阻害する要因となることが示されている．これらの結果は，固定効果を入れないモデル１でも，固定効果を入れたモデル２とモデル３でも，いずれも共通しており，係数にも大きな違いは見られない．なお，モデル１では，費用負担の 100 円／月の水準の係数が 3.767 であるから，係数が-0.447 の「静止画付き」は 11.9 円／月の効用（あるいはメリット）をもたらすと評価されているのに対して，ウェブサイトからの情報提供と比較して，ツイッターによる情報提供は 38.2 円／月，フェイスブックによる情報提供は 36.4 円／月のマイナスのメリット，すなわちデメリットとしてしか評価されていないという結果となっている．

モデル１に，性別と年齢の固定効果を組み込んだモデル２では，男性の係数が１％有意水準

表２　選択実験の係数の推定結果

項		モデル1	モデル2	モデル3
切片		-0.825 **	-0.685 **	-0.683 **
媒体	ツイッター	1.439 **	1.453 **	1.445 **
	フェイスブック	1.372 **	1.386 **	1.379 **
内容	静止画付き	-0.447 **	-0.461 **	-0.455 **
	動画付き	-0.038	-0.040	-0.037
費用	100円／月	3.767 **	3.804 **	3.791 **
	200円／月	4.043 **	4.082 **	4.071 **
性別	男性		-0.115 **	
年齢	20歳代		-0.671 **	
	30歳代		-0.331 **	
	50歳代		0.172 **	
	60歳代		0.153 **	
	70歳以上		-0.064	
職業	会社員・団体職員			-0.258 **
	公務員			-0.109
	自営業			-0.086
	パート・アルバイト			-0.073
	学生			-1.057 **
	無職			-0.014
	退職後年金生活			-0.098
	その他			-0.167
最大対数尤度		-12259.7	-12125.5	-12124.5
決定係数		0.4178	0.4241	0.4214

推定された係数は，（利用しない）／（利用する）の対数オッズに対するものであり，**は１％有意水準で，*は５％有意水準で係数が有意であることを示す．

Q19－1　もしも、身に覚えがないのにもかかわらず、ゴミ出しのルールに違反してゴミ収集日以外にゴミ出しをしたとして、あなたの名前が市区町村が情報発信をしている公式のフェイスブックのコメント欄に書き込まれ、1時間後には削除されましたが、1時間の間に誰とは特定できませんが、何人かの人々には閲覧されたとします。
このような状況を想定して、あなたはフェイスブックを管理している市区町村に損害賠償を求めますか。

○求めない ――→ Q19-3へ
○求める ――→ Q19-2へ
○わからない ――→ Q19-3へ

Q19－2　（損害賠償を求めると回答の方）

■ 市区町村に求める損害賠償の金額はいくらですか。

○1万円以上
　→ ○5万円以上 ○5万円未満 → 具体的な賠償補償の金額は［　　万円
○1万円未満
　→ ○5千円以上 ○5千円未満 → 具体的な賠償補償の金額は［　　千円

図６　中傷被害への賠償額を問う質問

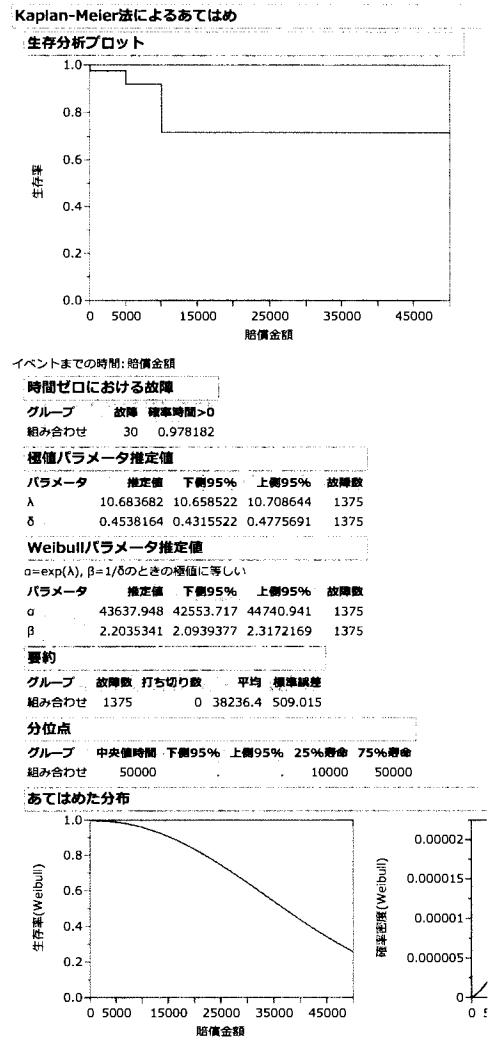


図７　生存時間分析による WTA 曲線のあてはめ

でマイナスで有意である．これは，女性よりは男性の方がインターネットやソーシャルメディアを通じた情報提供を利用するという傾向を示すものである．年齢については，70 歳以上を除いて推定された係数は１％有意水準で有意であり，40 歳代を基準にすると，40 歳未満の若い世代はデジタルによる情報提供を享受しやすく，50 歳代以上はデジタル媒体を好まない傾向が示されている．また，モデル３は，性別と年齢の代わりに職業を固定効果として組み込んだモデルであるが，家事専業を基準とすると，会社員・団体職員と学生の係数が１％有意水準で有意でマイナスであることから，職場や学校で PC やデジタル端末に慣れ親しんでいればデジタル媒体での情報取得への抵抗が小さいことが示唆されていると考えられるが，詳細な分析が必要である．

以上から，インターネットならびにソーシャルメディアを通じての自治体からの情報提供に関しては，静止画付きの情報提供はプラスに評価されるが，現時点では，フェイスブックやツイッターによる情報提供は，即時性や公開性や双方向性といったメリットがあると言われてはいるものの，多くの住民にはそのようなメリットは評価されておらず，逆にマイナスの評価となっているという，我々の予想とは異なる結果が得られている．

一方，ソーシャルメディアを通じての自治体からの情報提供に関するデメリットの定量評価については，情報漏洩や風評被害が発生した場合の補償額としての支払い限度ならびに被害者となった場合の受け取り希望額を，それぞれ CVM の WTP(Willingness to Pay)と WTA (Willingness to Accept) を聞くことで推計を試みた．

具体的には，ソーシャルメディアへの書き込みによる軽微な中傷による名誉棄損に相当する仮想の状況に対して，損害賠償を求めるかどうか，損賠賠償を求めるとすればいくらを求めるかを，２段階２項選択法を適用して設問し，図６に示す質問を回答者に示して回答を得た．自治体への損害賠償を求めるとの回答は，28.2%の 1,409 サンプルで，そのうちで賠償金額に関する２段階２項選択法による質問に回答している 1,375

サンプルの回答データに生存時間分析を適用して自発的受入額（WTA：Willingness To Accept）曲線を推計した結果を図7に示している．図7から，賠償額の平均は38,236.4円と推計されており，この金額は決して小さい金額とは言えないことが分かる．

4．おわりに

本稿では，2013年度末に全国を対象に実施した5,000人規模の「自治体からのソーシャルメディアによる情報発信に関するウェブ調査」の回答データをJMPで解析することから，一般の住民と比べると比較的にインターネットやソーシャルメディアの利用頻度が高いと考えられるセグメントの自治体から提供される地域や生活に関わる情報の利用状況と明らかにし，選択実験によるソーシャルメディアのメリットの定量評価とCVMによるソーシャルメディアのデメリットの定量評価を試みた．その結果，デメリットについては調査前の仮説に沿う結果が得られたが，ソーシャルメディアの即時性や双方向性や公開性といったメリットについては，現時点ではメリットとは認識されず，ウェブサイトからの情報提供よりもソーシャルメディアによる情報提供は低い評価しかされていないという結果が得られた．ただし，これらの結果は予備的分析の段階での結果であり，今後も，回答者のソーシャルメディアの利用状況などを加味した分析を行うとともに，ソーシャルメディアを活用して広報・公聴活動を行っている先進自治体を訪問して調査結果を提示してのヒアリング調査を行い，自治体からの反応も踏まえて，研究を深化させていきたいと考えている．

謝辞

本研究は，公益財団法人電気通信普及財団の平成25年度の研究調査助成の助成金が授与されて行った「自治体からのソーシャルメディアによる情報発信の効果の定量的評価に関する研究」の一部である．ウェブ調査に協力いただいた回答者の方々に感謝いたします．

参考文献

有馬昌宏・加藤優希・藤田昌弘，「コンジョイント分析による官民連携地域ポータルサイトの住民評価」，『2009 ＳＡＳユーザー総会アカデミア／テクノロジー＆ソリューションセッション論文集』，pp.231-242，2009

財団法人地方自治情報センター，『フェイスブックやツイッター等を活用した行政サービスの充実について』，2013

庄司昌彦，「人間関係ベースの消費　－地域SNSが地域生活化に貢献するために」，『智場』，No.114，pp.122-126，2009

総務省，『平成25年版情報通信白書』（http://www.soumu.go.jp/johotsusintokei/whitepaper/h25.html），2013

谷本晴樹，「自治体はfacebookをどう使うべきか－武雄市の事例から考える」，http://www.grand-design.org/blog/ict/cat17/6facebook.htm，2011

谷本晴樹，「自治体の SNS 利活用　－国内外の「先進」事例に見る課題と展望」，ICT 将来政策研究会報告（http://www.slideboom.com/presentations/513185/），2013

土木学会土木計画学研究委員会，『非集計行動モデルの理論と実際』，土木学会，1995

藤尾俊幸・繆青・黒田佳代・田中有紀・川向肇・有馬昌宏，「ＪＭＰを利用したＣＶＭによる政策評価」，『SAS Forum ユーザー会学術総会2005論文集』，pp.415-424，2005

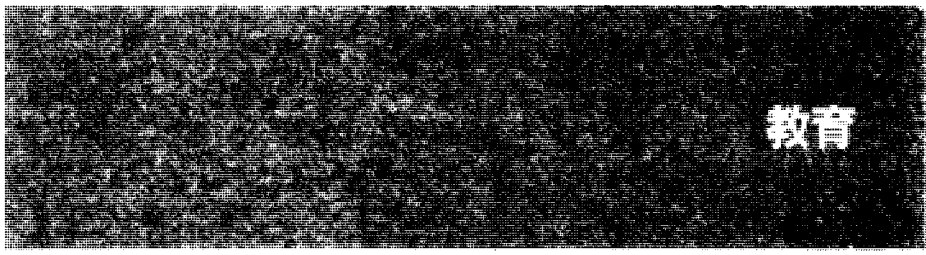
最近の公的統計データの利用について ～二次的利用の取組と統計のオープンデータの高度化を中心に～

高部 勲
総務省統計局統計調査部経済基本構造統計課 課長補佐

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

・ミクロデータの二次的利用に関する制度の概要
・匿名データ及びオーダーメード集計の概要・利用実績
・統計のオープンデータの高度化の概要
・第Ⅱ期基本計画における関連施策の概要
・公的統計データの高度利用に関する今後の動向


教育

大学と企業における統計教育とSAS

山之内 直樹
第一三共株式会社 データサイエンス部 主幹

堺 伸也

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

SASユーザ会世話人の我々は、医薬品業界のユーザとして企業が求める統計解析担当者として必要な知識と経験、素養について、アカデミアの指導教官、学生と懇談した。製薬企業の研究開発業務の統計とSASの有用性を紹介するとともに、大学と企業の統計教育とSASについての情報交換も行った。まず2012年は、医薬品メーカー、CRO(Contract Research Organization)と統計教育及びSASの有用性について各大学の統計教育の担当教官と交流会を開催した。さらに2013年は、医薬品、CRO15社で実施したアンケートから企業内の教育を紹介した。これらの内容を通じて、企業が大学に求めるべき統計家のコンピテンスとしてどのような教育が必要となるのか、そしてSASをどのように広げていくか等の議論が必要があると考え、全国の大学で、統計学の指導教官、学生と座談会を行い、医薬品業界で必要とされている統計学的、SASの知識、スキル、また企業が求める人材要件等について交流した。今回はその内容を報告し、統計教育とSASの議論を行う。

科学的マーケティング手法による
大学マネジメント・サイクルの持続的発展
— 山形大学EM部の「学生を知り抜く」IRへの挑戦 —

福島真司
山形大学エンロールメント・マネジメント部

Sustainable Development of University Management Cycle with the Scientific Marketing Approaches
- A Challenge of the Institutional Research of "Learning Students' themselves"
at Department of Enrollment Management, Yamagata University –

Shinji FUKUSHIMA
Enrollment Management Department / Yamagata University

要旨

エンロールメント・マネジメントは、大学の持続的発展を目的とする、科学的マーケティングを用いた大学マネジメント手法のことである。大学におけるマーケティングとは、学生募集や寄付募集等だけをさすのではない。私たちの学生の価値を創造し、その価値の最大を実現し続けるための、組織一体となった活動が、大学におけるマーケティングである。

大学マーケティングを実現するためには、学生を知り抜くための Institutional Research （以下、IR）が欠かせない。入学前の接触情報から、入学時点での期待感や満足度、入学後の GPA[1] 、学生生活や教育サービスへの満足度、授業評価のスコア、目標の達成感、課外活動や奨学金の履歴、就職、卒業後の大学教育への満足度や要望等を調査、分析することなしには、私たちの学生を知り抜くことはできない。

山形大学では、入学前から卒業後までの情報を一貫して分析するための IT インフラである「総合的学生情報データ分析システム」の構築に 2010 年度より着手し、真のマーケティング志向の組織運営をめざし、IR に取り組んでいる。その要諦は、組織文化の醸成に他ならない。

本発表は、山形大学における IR の取り組みについて、そのコンセプトと業務の実際を報告し、組織文化の醸成に、ビジュアル・アナリティクスがどのように寄与しているかの実践的事例を報告するものである。

キーワード：エンロールメント・マネジメント、マーケティング、IR、総合的学生情報データ分析システム

[1] Grade Point Average：合格した科目の成績の評定を S,A,B,C の 4 段階で行い、その成績を平均化したもの。

1　はじめに

近年、日本の大学においても、エンロールメント・マネジメント（以下、EM）の重要性が説かれているが、その定義は、どのようなものであろうか。

EMは、1970年代に米国ボストン・カレッジに、アドミッション・セクションのディレクターとして従事し、母校である当該大学の危機的状況をターン・アラウンドさせたジョン・マグワイヤ博士が、最初に提唱したと言われる。大学の経営的な危機は、学生数に見合わない過剰な投資と、何より、学生数の減少による収入の減少が最も大きな要因となる。

学生募集において、多岐に亘るプロモーションや奨学金に大きな投資することで学生をかき集めても、退学や転学により学生を失えば、学生募集にかけたコストは無駄になる。また、教育内容に関しても、ただ高度なレベルの学問を提供しても、コースに集まった学生とアンマッチした内容であれば、不満につながる。大学が提供する諸教育サービスに対し、不満を持つ者が多くなれば、退学・転学者数の増加による収入の減少や大学の評判の低下による大学ブランドの毀損につながる。EMはこのような背景の中で考えられたメネジメント手法である。

2．EMとは何か

1970年代のボストン・カレッジは、「カソリックのハーバード大学をめざしていた」と揶揄される状態にあった。教授たちは、最優秀な学生だけを教えたがり、高いレベルの研究をすれば入学者は自然と集まってくると考えられ、当時の社会的な評価とアンマッチした学生募集戦略や教育内容の提供をし続け、入学者数の減少、退学者数の増加、寄付金の減少、さらには社会的な評価の下落という負のスパイラルに陥っていた。マグワイヤ博士は、その状態をマーケティング調査により、わかりやすく可視化し、ボストン・カレッジの採るべき戦略を明確にし、大学全体を改善するという意志決定へとつなげた。

この過程の中で、マグワイヤ博士は、大学の永続的な発展には、組織一体となったダイナミックなマネジメント・サイクルの重要性を認識する。上記の通り学生をかき集めても、不満足な学生は退学してしまうため、その場合、様々なコストが無駄になる。そこで、一旦、ボストン・カレッジに興味を持った学生を、効果的なプロモーションにより出願までつなげ、入学許可を出した学生を奨学金の最適な配分により入学までつなげる。入学した学生に対し、その学生の期待に見合った教育サービスを提供することで、満足した教育的な経験や学生生活を提供し、卒業後の社会で希望の職業や満足感のある生活を得られるように価値を付加する。卒業生となった後にも、大学との継続した良好な関係を築き、卒業生の子や孫の入学につなげたり、大学に対する遺贈も含めた寄付募集につなげるまでをマネジメントする。この個々の学生との一生涯のつながりを、大学が組織一体となって、科学的なマーケティング調査をもとにサイクルさせ続けるマネジメント手法が、エンロールメント・マネジメントである。

日本では、EMの定義を「学生支援策」や「教育の考え方」等とする向きもあるが、EMは、科学的マーケティング手法を用いた大学マネジメント手法のことである。当然ながら、ここで言うマーケティングとは、売り込みの術をさす狭義のマーケティングではなく、顧客価値の創造やその最大化を、組織一体となって実現するための活動をさす。大学におけるマーケティングセクションは、学生募集や寄付募集、学生担当窓口担当部署だけをさすのではなく、総務や人事、会計部署などのバックヤードの部署も含めた組織一体となった活動が大学マーケティングである。

2．IRの果たす役割

科学的マーケティング手法には、大学自体に関する調査分析が欠かせない。学生に関する調査もここに含まれるが、個人的な考えや憶測をベースに学生の実態を掴んだつもりで、大学マネジメントを行うことは、科学的な態度とは言えない。ここで必要となる大学自体に関する調査分析は、IRと呼ばれる。

ボストン・カレッジを例に挙げると、ボストン・カレッジに興味を持ってパンフレット等の資料を請求した者、オープンハウス（日本でのオープンキャンパス）に来場した者等のプロファイルを精緻に行い、その後、実際に出願した者と出願しなかった者をセグメントに分けて分析し、なぜ出願し、なぜ出願しなかったのかの要因を探る。また、入学許可を受け取った者のうち、実際に入学した者と、入学しなかった者をセグメントし、なぜボストン・カレッジを選び、なぜ他の大学を選んだのかの要因を探る。この調査結果を、わかりやすく可視化し、必要な学内の会議で共有し、学生募集戦略の策定、奨学金の設計、教育プログラムの改善等のプランニングに活かす。

学生募集に関する調査項目では、SAT等の大学進学適性試験の成績、出身高校のタイプ、学生の自宅と大学までの距離、大学寮への入寮希望、人種、保護者の支払い能力と奨学金の希望内容の詳細、学問的な志向、サークル活動やボランティア活動の経験等の多岐に亘る要素が対象となる。大学での教育的な経験への期待を把握することは大切である。同様に、入学前の期待と入学した後の満足が見合っているかを確認するため、学生満足度調査は重要であり、アンケートだけではなく、インタビューも加えながら実施される。

入学者数を増加させることだけで、大学の収入状況が改善され、経営的な危機が回避されるわけではない。米国の大学では、日本の大学に比較して転学のためのシステムが整備されており、入学した大学が期待に見合わないものであった場合、学生は他の大学へと転学する。転学を含めた退学を阻止し、在籍率を向上させることも、入学者獲得と同様に、大学経営上は重要である。不満足を理由に退学していく学生から、情報を得ることには困難が伴うが、転学を支援する部署において、当該学生が転学に至った経緯を情報収集し、進路先に関する情報を提供し、転学を成功させるためのサービスを提供することで、当該学生から情報を得る退学率を下げるために転学の支援をするという、一見矛盾を孕んだ取り組みの効果は大きかったとのことである。このようにボストン・カレッジでは、学生に関する多岐に亘る情報を収集し、分析した結果を、大学マネジメントに活かしている

膨大なデータ量を扱うIRには、データベースやデータ分析のITインフラの構築が欠かせない。IRに係るデータは、入学前の接触情報から、入学時点での期待感や満足度、入学後のGPA、学生生活や教育サービスへの満足度、授業評価のスコア、目標の達成感、課外活動や奨学金の履歴、就職、卒業後の大学教育への満足度や要望等多岐に亘っており、しかも大学が継続する限り、データ量は増え続けることとなる。一定以上の学生の数を持つ大学が、学生の様々な情報をクロスして分析し、その変化を経年的に見るような場合、手作業による分析は、時間コストが掛かり過ぎ現実的とは言えない。クライアントPC上で動くアプリケーションだけでは、データの格納、分析、結果の共有、セキュリティへの配慮、閲覧権限の制限等の要求が満たされない場合も出てくる。

3．私たちの学生を知り抜くこと

米国と、日本の高等教育を取り巻く環境は異なっており、マネジメント手法も、それぞれの環境に適合

する形で発展してきた。すなわち、米国のEMのあり方を、そのまま日本に適用することに必然性はない。山形大学EM部[2]は、2006年7月1日に、日本の大学初のEM専任部署として発足して以来、マーケティング志向のセクション運営を行ってきた。

山形大学EM部は、学生の大学入学前は学生募集活動を、入学後は学生満足度調査を始め様々な学生に関するIRの諸活動を、卒業後はニーズ調査を始め卒業生の大学へのロイヤリティ向上のための諸活動に関する業務に責任をもつ。設置経緯には、学生募集の危機的な状況が関係していたため、学生募集活動に業務の中心があるが、近年は、IRの諸活動に業務の中心を移しつつある。

2010年度から3年間の期間で採択された文部科学省概算要求事業「学生の大学への期待、満足度、成長の自覚、目標達成感等を向上させることを中心においた教育改革マネジメント・サイクルの実現」によって、「総合的学生情報データ分析システム」を整備し、IRにより一層重視する体制が整いつつあるが、コンセプトは「私たちの学生を知り抜くこと」である。学生の入学から卒業後までを、一連のデータとして保有し、様々な切り口で分析し、常に見える化し、トラックし続け、その結果を元に改善する仕組みをマネジメントのフローにビルトインすることで、大学は永続的な発展を遂げる可能性が高まる。

山形大学EM部が扱うデータは、学生の出身高校や受験前の大学との接触データ、入試に関するデータ、履修状況やICカード型学生証による出欠状況、成績、授業評価、就職状況、入学前・在学中・卒業後の満足度アンケート等の各種アンケートのデータと多岐に亘っている。これらの学生データを利用し、各学部等が必要とする分析を実施し、レポートを行うサービスを各学部に対して提供している。当初構築したシステムでは、データ・プレゼンテーションのためのダッシュボードやテンプレートの作成に対し、限られたシステム担当者やITベンダー企業のSE等のマニュアル作業に依存していたため、時間コスト、経費的なコスト共に、大きな負担があり、即時的な分析やレポート提供が困難であるという課題を抱えていた。

SAS Visual Analyticsの導入によって、ITベンダー企業のSEへの依頼がなくとも、データ・プレゼンテーションを行うことのできる環境が整い、時間コスト、経費的なコストを大幅に縮減することができることとなった。また、容量の大きなデータも短時間に分析可能とするインメモリ・アナリティクス機能によりデータのメモリ上への読み込みや処理の高速化が実現される。分析結果はipadなどでも閲覧可能であり、ユーザビリティがよく、素早く、美しく見られるビジュアルによって、これまで使用感や操作性の悪さから、使用を忌避されていた分析システムへの期待感が高まり、従前よりも、IRの学内認知が高まった。

IRは、特定の専門家が、専門家以外には解釈が難解な統計を駆使して行うものはなく、データ分析を必要とする部署の担当者が、自分自身の手で、簡易に、操作性よくデータを扱い、わかりやすく、美しい分析画面を見ながら、諸業務の評価や改善策の立案、意思決定に役立てることにその要諦がある。これがないと、IRが組織文化に浸透することは難しい。システム担当者にしか使いこなせないシステムや、難解でわかりにくい分析画面では、データ分析結果を閲覧しようとするモチベーションにも影響を与えるからである。IRの目的は分析自身にあるのではなく、データ分析を意志決定の業務フローのビルトインし、科学的な調査結果に基づいて、マネジメントのPDCAサイクルを回すことにあるため、システムの仕様にも、誰もが使いたくなるシステムであることが要求される。山形大学の従来のシステムには、柔軟性や拡張性にも課題があり、常に時間的・経費的なコストが膨らみ、各学部等の分析要望にも即時的に応えることができず、IRの浸透に困難があったが、SAS Visual Analyticsの導入によって、諸業務の意志決定や事後の評価がこれまでよりも支援され、EMの質の向上が期待できる可能性が高まった。

---

[2] EM部の設置当時は、EM室であった。その後、部に昇格した。

4．おわりに

マーケティングの潮流にも変化が見られる。マーケティング手法にフレームワークはあっても統一手法はないため、個々の大学が個々の大学のやり方を、トライ・アンド・エラーで見つけながら、個々の大学の目的に叶った業務活動での精度を上げる続ける必要がある。常に、私たちの学生を知り抜くことが、そして、社会の期待を知り抜くことが、学生の価値を創造し、その価値を最大化し続けることに直結する。EM を支える IR では、データという FACT をもとに意志決定を行うという組織文化の醸成が重要となる。IT インフラが大学マーケティング、大学マネジメントに与える影響は大きい。

参考文献等

John Maguire「To the organized, go for students」『Boston College Bridge Magazine』Boston College（1976 年）

ダニエル・ピンク、大前 研一（(翻訳)『ハイ・コンセプト「新しいこと」を考え出す人の時代』三笠書房（2006 年）

John Maguire、Lawrence Butler、Maguire Associates『Em=C2: A New Formula for Enrollment Management』Transworld Pub（2008 年）

フィリップ・コトラー『コトラーのマーケティング 3.0　ソーシャル・メディア時代の新法則』朝日新聞出版（2010 年）

"エンロールメント・マネジメントの国内スタンダード" 創造に向けた新しい情報基盤の構築へ。 400 項目を超える要求に応える最適解として SQL Server 2008 を中核とするソリューションを採用

http://www.microsoft.com/ja-jp/casestudies/yamagata-u.aspx（2011 年 9 月掲載）

山形大学、SAS® Visual Analytics の導入でより的確な意思決定とエンロールメント・マネジメントを加速～直感的ユーザーインターフェイスで各学部担当者による即時的データ抽出と柔軟なレポーティングを可能に～

http://www.sas.com/offices/asiapacific/japan/news/press/201312/19.html（2013 年 12 月掲載）

マーケティング管理

インターネット時代の医薬品営業に関する考察

武藤　猛
MarkeTech Consulting 代表

Pharmaceutical Sales in Internet Era

Takeshi Muto
President, MarkeTech Consulting

要旨

患者、医療従事者、医療機関、製薬企業など、医療に関連したあらゆる分野でインターネットが活用される時代となり、医療用医薬品の情報に関しても、医師が積極的にインターネットから直接情報を入手し、活用することがごく普通となっている。このため、製薬企業にとって、従来の MR（医薬情報担当者）に依存した営業戦略の見直しが迫られている。本論文では、数種類の医薬品について、医師に対して MR およびインターネットで情報提供した場合の効果を定量的に分析した。目的変数を各医師の処方量（医薬品売上高）、説明変数を MR またはインターネットによる情報提供回数とし、これらの交差項を含め、重回帰分析を行なった。その結果、いずれの医薬品についても、MR に加えてインターネットによる情報提供は売上高を押し上げる効果があることが分かった。一方、交差項は I 種類の医薬品についてのみ有効であった。このような交差項（つまり MR とインターネットの相乗効果）が有効な条件について考察を行った。限られた事例ではあるが、今回の分析結果を踏まえ、医薬品営業において、インターネットを効果的に活用する条件について検討した。

キーワード：医薬品営業、インターネットによる情報提供、効果測定

1．医薬品マーケティングとインターネット

1.1　医薬品マーケティングにおけるインターネット活用の現況

患者、医療従事者、医療機関、製薬企業など、医療に関連したあらゆる分野でインターネットが活用される時代となった。医師に対する医療用医薬品の情報に関しても、医師が積極的にインターネットから直接情報を入手し、活用することが日常的に行われている。このため、製薬企業にとって、従来の人的方法、つまり MR（医薬情報担当者）に依存した営業戦略も見直しを迫られている。また、国民医療費の適正化のために、製薬企業の MR に依存した高営業コストに対して厳しい眼が向けられつつあることも、製薬企業におけるインターネット活用の方向を加速している。

ある調査によれば、医師のインターネット利用時間は一日平均 2.9 時間である(1)。別の調査によれば、医師が情報収集に費やす時間は、インターネットと MR がほぼ同じである(2)。これらの調査結果を総合すると、医師の視点からも、また製薬企業の視点からも、医療用医薬品の情報に関するインターネットの活用が不可欠な時代となっている。

1.2 医薬品マーケティングにおけるインターネットの役割

インターネットによる医師への医療用医薬品に関する情報提供（「eディテーリング」あるいは「eプロモーション」などと呼ばれるが、本論文では前者を用いる）も商業ベースで盛んになった。医薬品は、適切な処方をしないと（あるいは適切に処方しても）患者に予期しない副作用をもたらす可能性がある。一般消費財と異なり、この点が医療用医薬品の場合に、MR から医師への直接情報提供（ディテーリング）が欠かせない理由である。このディテーリングにおいては、単に当該製品の特長や適用対象患者、適正処方量だけでなく、安全性に関する情報の提供が欠かせない。従って、医療用医薬品の場合は、すべての情報提供をインターネットだけで行うことは現時点では考えにくく、製薬企業にとって MR チャネルとインターネットチャネルとを使い分けあるいは併用して相乗効果を考える必要が生じてくる。

MR と e ディテーリングとを比較すると、前者は効果（処方への影響度）が高いが、効率（情報提供1回当りコスト）が低く、一方 e ディテーリングは、効率は高いが効果は低い、というのが「通説」である。この「通説」は、各種情報の情報源と処方への影響度を調べた調査結果からもある程度は推察される[3]。そこで、インターネットによる情報提供と MR による情報提供が共存するという現実的な条件（「MR＋e」と呼ばれる）下で、両者の効果を定量的に検証することが重要となるが、そのような研究はほとんど公開されていない。なお、医師の医薬品情報入手時の行動を調査した結果によると、医師は、インターネットからの情報を MR に確認し、MR の情報をインターネットで確認することが判明している[4]。このため、「MR＋e」の効果測定には、両者の相乗効果の有無を考慮することが重要となる。

1.3 本論文の目的

本論文の目的は、医師に対して MR およびインターネットで情報提供した場合の効果（処方量、つまり医薬品売上）を定量的に分析することである。医師に対して、MR およびインターネットで情報提供した場合の各々の効果を評価するとともに、両者の相乗効果が存在するかどうかについても検討する。

2．各情報チャネルの効果測定の考え方

医薬品マーケティングにおいて、製薬企業から医師への情報提供チャネルには、MR だけでなく、講演会や研究会、MS（医薬品卸企業のマーケティングスタッフ）、製薬企業が設置している医療関係者向けコールセンター、e ディテーリング、および DTC（Direct To Customer の略で、新聞やテレビにおける疾患啓蒙広告を意味している。これを見た潜在患者が医師を受診する可能性があるので、製薬企業から医師への間接的な情報提供チャネルと考えられる）などがある。これらの情報提供チャネルが、「効果対効率」のトレードオフ曲線上に位置していると考えられる。なお、このトレードオフ曲線は、「効果×効率＝1 処方当り獲得コスト（ROI）＝一定」の曲線である。ただし、データが限られている現状では、このトレードオフ曲線はあくまで作業仮説に過ぎない。

各情報チャネルの効果測定の考え方は、次の通りである。

①各情報チャネル別の訪問・参加・アクセス回数を医師単位に記録し、データベース化する

②一定期間（半年・1年）のデータを用いて、目的変数＝処方量（売上高）、説明変数＝情報チャネル別訪問・参加・アクセス回数として、回帰分析する

③各情報チャネルの偏回帰係数から、各情報チャネルの効果を評価する

データベースのイメージを図表1に示す。

図表1 各情報チャネルの効果測定のためのデータベース

医師_ID	情報チャネル別訪問・参加・アクセス回数(2013年4月～9月)						
	MR	MS（卸営業）	コールセンター	講演会	学会セミナー	eディテーリング	製薬企業ウェブサイト
1000001	20	0	0	1	1	0	0
1000002	15	2	0	2	0	1	3
1000003	0	3	3	0	1	3	2
1000004	20	0	0	1	1	0	2
1000005	13	4	0	0	0	1	0
1000006	8	3	1	1	1	2	1
1000007	20	0	0	2	1	0	0
1000008	6	2	2	0	0	3	1
1000009	15	3	0	1	0	1	0
1000010	20	0	0	0	1	0	2
…	…	…	…	…	…	…	…

このような各情報チャネルの効果測定用統合的データベースを構築している製薬企業はまだ少数と考えられる。本論文では、MR とｅディテーリング（いわゆる「MR＋e」）に限定して、図表１に相当するデータベースを構築し、効果測定を行った結果を報告する。

３．「MR＋e」の効果測定例

3.1　効果測定の方法

分析対象とするｅディテーリング（「eDTL」と略称する場合がある）は、5 種類の薬剤について、登録された医師に対して、一定期間行われたものである。5 種類のうち 4 種類はマス領域（広範囲の診療科の医師が処方する薬剤）、1 種類は非マス領域（特定の診療科の医師が処方する薬剤）である。ｅディテーリング実施時において、各薬剤は、上市後かなりの年数が経っている。なお、以下に示す効果測定結果は、実施例を参考にして作成した、あくまで仮想的なもので、実施例そのものとは異なっている（ただし、結論には影響しない）。

効果測定の方法は次の通りである。目的変数は、ｅディテーリング実施期間中の売上高（処方量と同じ。元の分布は偏りが大きいので５段階の順序変数に変換）、説明変数は、同期間中の MR 訪問回数およびｅディテーリング視聴回数である。効果測定のための重回帰式（交差項を含む）は次の通りである。

$$y = \beta_0 + \beta_{MR}x_{MR} + \beta_e x_e + \beta_{MR\times e}x_{MR}x_e + \varepsilon$$

ここで、

y ：売上高

x_{MR}, x_e ：MR 訪問回数および eDTL 視聴回数

β_0, β_{MR}, β_e, $\beta_{MR\times e}$ ：パラメータ（偏回帰係数）

ϵ ：誤差

3.2　効果測定結果

5 種類の薬剤に対する「MR＋e」の効果測定結果を図表 2 に示す。ただし、表中の偏回帰係数は、各回帰式において、MR の効果を表わす偏回帰係数を 1.0 とした場合の、e ディテーリングと「MR×e ディテーリング」の交差項の偏回帰係数を示している。

図表2 5種類の薬剤に対する「MR＋e」の効果測定結果

対象製品	データ件数	重回帰分析*				
		自由度調整 R2乗	分散分析 (F値)	分散分析 (p値)	偏回帰係数 (eDTL) β_e/β_{MR}	偏回帰係数 (MR×eDTL) $\beta_{MR\times e}/\beta_{MR}$
A	45,000	0.0069	130.4	<.0001	3.416	0
B	5,000	0.5951	1933.9	<.0001	0.177	0.275
C	30,000	0.0279	419.7	<.0001	2.548	0
D	45,000	0.0371	935.3	<.0001	2.182	0
E	40,000	0.0111	217.7	<.0001	5.965	0
(平均)	33,000	-	-	-	2.858	-

*) 偏回帰係数は、MR訪問回数の偏回帰係数に対する比率を表示している; 偏回帰係数はすべて有意である(p<0.05)

図表 2 に示す重回帰分析結果において、モデル全体および偏回帰係数はすべて有意であった。ただし、「MR×e ディテーリング」の交差項は、非マス領域の製品 B だけが正の値で、他はすべて 0 である。この理由については後述するとして、製品 B について、重回帰式を記す：

y（製品 B 売上高）$=\beta_0$　＋1.0×（MR 訪問回数）
＋0.177×（e ディテーリング視聴回数）
＋0.275（MR 訪問回数）×（e ディテーリング視聴回数）

製品 B についてのみ、「MR＋e」の相乗効果が観測された理由をまとめる：

①専門医が処方する製品（非マス領域製品）であり、あらかじめ「MR＋e」の戦略を検討した

②専門医も満足する、専門性の高いコンテンツを開発した

③対象医師の中で、 e ディテーリング視聴率が比較的高かった（約 25%）

④対象医師の約 60%は MR が訪問し、フォローした

⑤上記のように、MR 訪問と e ディテーリングを効果的に関連付けたことが相乗効果に繋がった

⑥以上の結果として、平均売上高アップなどの効果が得られた

3.3　「MR＋e」の通年効果の推定とその意味

e ディテーリングは通常、キャンペーンとして行われ、期間は限定される。そこで、「MR＋e」の通年効果の推定を行う。そのため、e ディテーリングは 3 か月間実施されたと想定し、それ以外の期間は、e ディテー

リングの効果はゼロと仮定する。このような想定の下で、図表 2 の結果をマス領域と非マス領域に分けて、年間の平均売上高に及ぼす効果を推定したのが図表 3 である。表中の「平均売上高比率」は、MR のみが訪問した医師の処方量（売上高）に対する、MR 訪問に加えて e ディテーリングも視聴した医師の処方量（売上高）の比率である。

図表3「MR＋e」の通年効果の推定

期間	薬効領域の分類	平均売上高比率:（MR+eDTL）／（MR）	備考
eDTL期間	マス領域平均	1.48	製品B以外
	非マス領域平均	3.82	製品B
1年間	マス領域平均	**1.12**	1+(1.48−1)÷4
	非マス領域平均	**1.71**	1+(3.82−1)÷4

図表 3 によれば、年間の効果は非マス領域製品の方がマス領域に比べて大きい。この結果は、従来「医薬品マーケティングにおいて、インターネットはマス領域製品に対して効果がある」とされる「通説」を否定するものである。

このような、マス領域製品と非マス領域製品に対する効果の違いは、医師ターゲティングの考え方から解釈することが可能である。医師ターゲティングとは、販売ポテンシャル（患者数）の高い医師のセグメント（HP=High Potential セグメント）にディテーリングを集中することで売上への効果が高まる、とするマーケティングの基本的な考え方である[5]。大勢の患者で極めて多忙な HP セグメントの医師が、時間を割いて e ディテーリングを視聴したくなるようなコンテンツと、視聴前後の MR によるフォローが不可欠である。この点に関して、マス領域製品の場合は、対象医師数が多いことから MR による密接なフォローが難しい。他方、非マス領域製品の場合は、対象医師数が限られ、元々MR がすべての医師に対して密接なフォローを行なっている（特に HP セグメントの医師に対して）ところにインターネットが情報提供チャネルとして追加されたため、「MR＋e」の相乗効果が得られたのではないかと考えられる。今回の効果検証では、非マス領域製品が唯 1 つだけなので、現段階ではあくまで作業仮説である。今後、e ディテーリングは、MR によるディテーリングと同様に、ターゲティングの観点から効果を検証する必要がある。

3.4　「MR＋e」の売上高への効果シミュレーション

かなり大胆ではあるが、3.3 の結果を用いて、「MR＋e」の効果が製品売上高に及ぼす効果をシミュレーションする。図表 4 の「MR＋e」の通年効果の指標と、ターゲティングとを組み合わせ、非マス領域の製品およびマス領域の製品の売上高への効果シミュレーションを行った結果を図表 5 および図表 6 に示す。なお、これらの表で、LP、MP、および HP は、販売ポテンシャル（当該疾患の患者数）に応じて医師を分類（セグメント化）したもので、LP=Low Potential、MP＝Middle Potential、および HP=High Potential を意味する。

図表4 マス領域における「MR＋e」の効果シミュレーション

・製品売上高：500億円

・対象医師数：90,000人

・MR訪問：HPセグメントの医師のみカバー

医師セグメント	LP	MP	HP	(計／平均)	備考
①医師数(人)	30,000	30,000	30,000	90,000	各セグメントに3分の1
②売上高(億円)	50	100	350	500	経験則による
③「MR＋e」の売上高アップ効果	6%	6%	12%	8%	MR訪問なしの売上高アップ率はHPの半分と想定
④eDTL視聴医師(比率)	4%	5%	6%	5%	セグメント毎に視聴率は異なると想定
⑤eDTL視聴医師(人)	1,200	1,500	1,800	4,500	①×④
⑥eDTLによる売上高増加(億円)	0.12	0.30	2.52	2.94	②×③×④
⑦eDTLによる売上高増加(比率)	0.2%	0.3%	0.7%	0.59%	⑥÷②
⑧eDTLのコスト(億円)				0.45	推定
⑨eDTLのROI(倍)				6.5	⑥÷⑧

図表5 非マス領域における「MR＋e」の効果シミュレーション

・製品売上高：100億円

・対象医師数：12,000人

・MR訪問：LP、MP、HPの全セグメントの医師カバー

医師セグメント	LP	MP	HP	(計／平均)	備考
①医師数(人)	4,000	4,000	4,000	12,000	各セグメントに3分の1
②売上高(億円)	10	20	70	100	経験則による
③「MR＋e」の売上高アップ効果	50%	60%	70%	60%	セグメント毎に売上高アップ率は異なると想定
④eDTL視聴医師(比率)	15%	20%	25%	20%	セグメント毎に視聴率は異なると想定
⑤eDTL視聴医師(人)	600	800	1,000	2,400	①×④
⑥eDTLによる売上高増加(億円)	0.75	2.40	12.25	15.4	②×③×④
⑦eDTLによる売上高増加(比率)	7.5%	12.0%	17.5%	15.4%	⑥÷②
⑧eDTLのコスト(億円)				0.48	推定
⑨eDTLのROI(倍)				32.1	⑥÷⑧

図表4に示したように、マス領域では、①視聴医師の比率が小さいこと、②MR訪問が可能なのはHPセグメントのみであり、「MR＋e」の相乗効果が限定的であること、の２つの理由で、全体売上高への効果は小さいと考えられる。一方、図表5によれば、非マス領域では、①視聴医師の比率が大きいこと、②全医師にMRが訪問し、全セグメントで「MR＋e」の相乗効果が期待できること、の２つの理由で、全体売上高への効果はかなり大きいと考えられる。

まとめると、多くの前提条件はあるが、マス領域製品に比べて、非マス領域製品の方がeディテーリングの売上高の押し上げ額が大きく、従ってeディテーリングの費用対投資効果（ROI）も大きいことが分かる。

3.5 医薬品市場全体に対する「MR＋e」の効果の評価

さらに大胆ではあるが、現在の医療用医薬品市場全体に、「MR＋e」の効果はどの程度あるのかを推定してみた。結果を図表6に示す。

図表6 医薬品市場全体に対する「MR＋e」の効果の評価

項目		値	備考
医薬品市場規模	①2012年（億円）	95,600	IMSジャパン発表（薬価ベース）
MR	②MR総数（人）	63,846	MR白書
	③MR生産性（億円／人）	1.50	①÷②
	④MR人件費（億円／人）	0.20	想定
	⑤MR総人件費（億円）	12,769	②×④
	⑥MRのROI（倍）	7.5	①÷⑤
	⑦年間ディテーリング回数（回／人）	1,000	「有効」ディテーリングは5回／日、稼働日は200日
	⑧年間総ディテーリング回数（回）	63,846,000	②×⑧
	⑨1ディテーリング当り売上高（万円／回）	15.0	①÷⑧
	⑨1ディテーリング当りコスト（万円／回）	2.0	⑤÷⑧
「e」	⑩医師向け「e」市場規模（億円）	200	「e」サービス企業の売上高から推定
	⑪a「e」の推定ROI（倍）－高め推定	7.5	「MR＋e」の効果測定例を参考にした推定（＝⑥）
	⑪b「e」の推定ROI（倍）－低め推定	3.7	MRのROI（⑥）の半分と推定
	⑫a「e」でもたらされた売上高（億円）－高め推定	1,497	⑩×⑪a
	⑫b「e」でもたらされた売上高（億円）－低め推定	749	⑩×⑪b
「e」の市場へのインパクト	⑬a市場全体への「e」のインパクト（%）－高め推定	1.6%	⑫a÷①
	⑬b市場全体への「e」のインパクト（%）－低め推定	0.8%	⑫b÷①

図表6に示したように、インターネットは、現時点で医薬品市場全体の1%程度に影響を与えていると考えられる。製薬企業によっては、売上高の数%程度に達している可能性がある。これは、特に専門領域で効果的なインターネット活用を行なっている製薬企業の場合である。

4．MR活動の質を高めるためのインターネット利用の考え方

4．1 売上アップの3要素から見た「MR＋e」の最適化

医薬品営業の観点から、売上高を決定する要素は、重要な順に、「ターゲティングの精度」、「MR活動の質」、および「ディテーリング回数」の3つである[6]。インターネットの効果もこの3要素に分けて考える必要がある。

まずターゲティングの重要性については、MRもインターネットも変わらない。ただし、インターネットの場合は、科学的な根拠からターゲティング医師が選定されているとは必ずしもいえないので、今後改善の余地がある。

次に、MR活動の質であるが、インターネットが医薬品マーケティングにおいて主に貢献できるのが、この点であると考えられる。つまり、MRから提供される製品情報や疾患、症例情報をインターネットで補い、医師の個別の医療ニーズに対応することで、結果としてMR活動の質を補完するということである。

最後に、ディテーリング回数については、最近MRに対する訪問規制を行なう医療機関が増え、MRが医師に面会しにくくなっているので、インターネットがMR訪問を部分的に代替できる可能性がある。ただし、eディテーリング視聴1回分が、MR訪問1回分と比べてどの程度の効果があるかについては、マス領域と非マス領域の違い、イコンテンツの質と更新頻度、上市後の年数、当該疾患に対する治療ニーズの変化など、影響する因子が多く、今後さらに研究する必要がある。

以上を、「MR＋e」の最適化として、図表7にまとめた。

図表7 売上アップの3要素から見た「MR＋e」の最適化

プロモーションの種類	売上アップの3要素：①ターゲティング	売上アップの3要素：②ディテーリングの質	売上アップの3要素：③ディテーリング回数	全体的評価
(A) MR単独	・MR活動の効果を最大化する訪問先 ・製薬企業側で設定可能 ・「3×3」マトリックスのHPセグメント	・教育訓練・OJTで維持・改善可能 ・MRアンケートや医師アンケートで測定	・MR活動の効果を最大化する訪問回数 ・MRが設定可能	・営業戦略として実施可能 ・効果が実証されている ・改善のためのツールが充実
(B) eディテーリング（eDTL）単独［MR活動との積極的連携なし］	・eDTL視聴医師は結果として決まる ・eDTL案内を行う医師は事前に決定できる ・精密なターゲティングは困難である	・視聴コンテンツの質で決まる ・専門／非専門医師向けは分ける必要 ・プッシュ型コンテンツには限界がある ・プル型コンテンツには魅力が必要	・ワンショットは3か月以内が多い ・継続するには、コンテンツを更新する ・リピート視聴には魅力が必要	・効果はあるが安定していない ・継続実施にはコストが掛かる
(C) 「MR＋eDTL」の積極的連携	・HPセグメント中心のMR訪問が基本 ・eDTL：MR訪問を補強 －ターゲティング戦略を阻害しない －「e」視聴医師比率を高める	・MRとeDTLで得意分野を棲み分ける －MR：症例、安全情報 －eDTL：疾患、薬剤に関する情報 ・アクセス分析でコンテンツ評価・改善	・MR：HPセグメント中心に訪問計画 －医師のニーズに合わせる ・eDTL：定期的に発信 －定期テクアクセスがあるよう更新	・「MR＋eDTL」の戦略が不可欠 ⇒「MR×eDTL」の相乗効果 ・MR主導で安定効果を期待

4．2　インターネット活用の落とし穴

最後に、医薬品マーケティングにおける「インターネット活用の落とし穴」について述べよう。医薬品マーケティングにおいて、インターネットは有用ではあるが、当然のことながら万能ではなく、特に下記の点に留意する必要がある。

①多忙な高ポテンシャル医師（HP セグメントの医師）はそもそもインターネットから情報入手する時間が限られており、優秀な MR から、短時間で要領よくまとめた情報を入手する方を望む可能性が高いこと（また、このような医師に対しては、各社とも MR が高頻度で訪問しているので、直接聞く機会が多い）

②感情のないインターネットだけでは、営業活動の基本である医師との信頼関係を構築することは困難であること

③e ディテーリングのコンテンツ視聴には意外と時間がかかるので、限られた医師のインターネット利用時間を獲得することは、MR の面会時間獲得以上に競争が激化する可能性が高いこと（MR の訪問回数を競う SOV=Share Of Voice の時代から、インターネットの視聴時間を競う SOI=Share Of Internet の時代へ？）

④ほとんどの製薬企業がインターネットで情報提供するようになると、他社より優れたコンテンツを継続的に提供するためには、コンテンツ開発のために、多額の費用が掛かる可能性があること

⑤「MR＋e」の効果を高めるには、「MR×e」、つまり相乗効果を発揮できるような薬効領域の製品（非マス領域）の製品を対象にする方が望ましく、どのような製品でも e ディテーリングが高い ROI を発揮するとは限らないこと

⑥「MR＋e」において、相乗効果を含めた効果を期待するならば、当面、「e」が MR を代替する（つまり、e ディテーリングで MR が不要になる）ということにはならない；逆に言えば、「e」の積極的活用は、現状の MR 人数を維持した上でコンテンツ開発費用が掛かるという具合に、製薬企業の営業コストのアップをもたらす可能性もあるので、ROI に十分留意する必要があること

5．まとめ

①インターネットによる情報提供（e ディテーリング）も、MR 活動と同様に、「ターゲティングの精度」、「MR 活動の質」、および「ディテーリング回数」の 3 要素で、各々の最適化を考えればよい

②5 種類の製品について、「MR＋e」の効果測定を行ったところ、いずれの製品でも「MR」および「e」の効果が認められたが、相乗効果「MR×e」は 1 種類の製品でのみ認められた

③上記の相乗効果は、非マス領域の製品に対するものであったが、この結果が一般的であるかどうかについては、今後の検証が必要である

④医薬品マーケティングにおけるインターネット活用は、慎重な戦略の下で実施することにより、売上高アップをもたらす；しかし、「MR×e」の相乗効果がないと安定した効果は期待できないので、当面はインターネットが MR に代替する可能性は小さいと考えられる

参考文献

(1) 社会情報サービス：医師メディア調査　S-DMR　2014 年版（2014 年）

(2) 熊西弘：医師多忙時代の医薬品情報源を探る、Monthly ミクス（2010 年 3 月号）

(3) 医療情報の情報源と処方影響度、Monthly ミクス（2012 年 11 月号）（オリジナルのデータは、シード・プランニング「医師による医療情報の入手動向 2012」）

(4) O2O で進化する「MR+e」モデル－医師の志向・行動に応じたチャネル戦略を、Monthly ミクス（2013 年 11 月号）（オリジナルのデータは、Medical Collective Intelligence Co., Ltd.）

(5) 武藤　猛：MR の生産性に関する考察、SAS ユーザ総会 2012（2012 年 8 月 2 日）

(6) 武藤　猛：MR の生産性アップと最適配置戦略、アンドテック社（2012 年）

マーケティングとデータ解析研究会

朝野熙彦　中央大学

Study Group of Marketing and Data Analysis

Hirohiko Asano, Chou University

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

マーケティング活動を効果的に実行するために、データ解析は貢献できているのだろうか？困難はどこにあって、どうしたら解決できるのかを検討しよう、という趣旨で研究会を開いた。昨年度の活動結果を報告する。

キーワード:因子分析、コーホート分析、ベイズ推定

G2：因子分析とクラスター分析

藤居誠
株式会社 東急エージェンシー

G2: A factor analysis and cluster analysis

Makoto Fujii
Tokyu Agency Inc.

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

アジェンダ

1. 利用の前提と実務上の対応
   - 実務的前提
   - 因子分析のフロー
   - 実務上行っていること
2. 因子分析の具体的手続き
   - ケース1
   - ケース2
3. 因子の回転
   - バリマックス回転
   - プロマックス回転

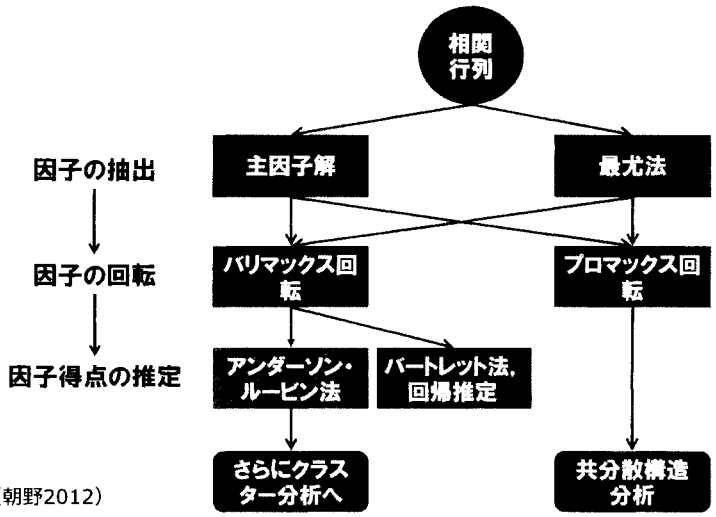
1. 利用の前提と実務上の対応

実務的前提

- なぜ因子分析なのか
  - 生活者の潜在意識を理解する
    - 生活者自身も意識していない価値観は、変数として直接観測することはできない
    - 生活者にとって回答可能な範囲の質問を設定し、その回答結果から、背後にある価値観を推定する必要がある
- なぜクラスター分析なのか
  - 因子分析だけでは、対象者のボリュームを把握しづらい
  - 「分ける」ことはしばしば「分かる」に通じる
    - 市場を理解し、市場に働きかけるため、管理しやすい数に市場を細分化することが効率的
    - 各クラスターと消費者属性とのクロス集計により市場の理解が促進される

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

因子分析のフロー



（朝野2012）

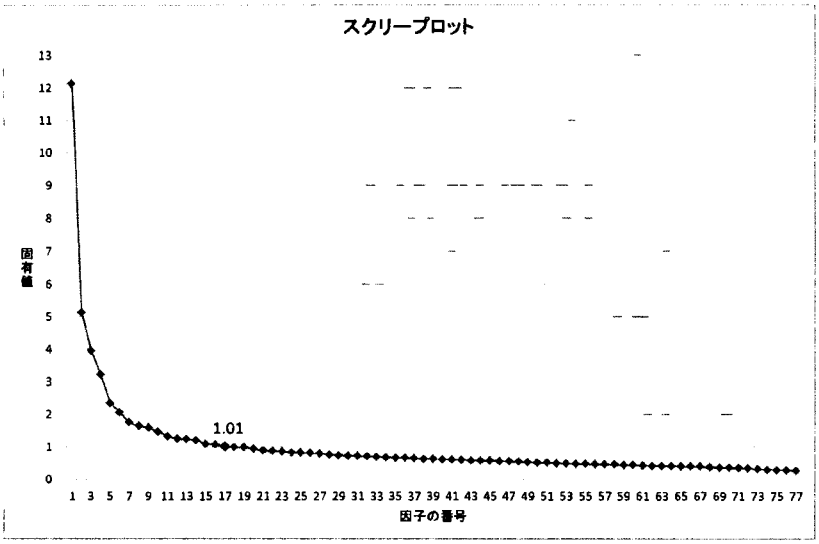
2. 因子分析の具体的手続き

- 分析データ概要
  - ネットリサーチ・パネル属性調査 ⇒ ライフスタイルクラスター
- ケース1：ごく普通に・・・
  - 因子分析
    - ステップ1：相関係数行列を分析対象にする
    - ステップ2：主因子解で因子抽出
    - ステップ3：因子の数は固有値＞1で決める
    - ステップ4：因子をバリマックス回転する
    - ステップ5：アンダーソン・ルービン法で因子得点を求める
  - クラスター分析
    - クラスター間の分離評価
- ケース2：変数をスクリーニングして・・・
  - 因子分析
  - クラスター分析

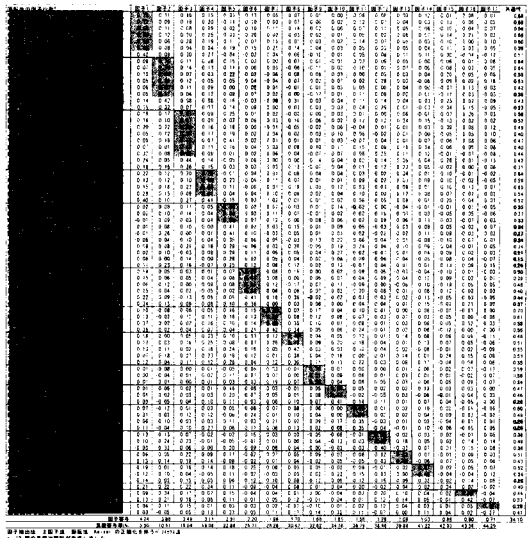
データ概要

- ネットリサーチ・パネル属性調査
  - 調査地域　東名阪エリア
  - 調査対象者　主婦
  - 回収標本数　5,629人
  - 調査方法　インターネット調査
  - 調査時期　2012年6月
- 分析データ
  - ネットリサーチ・パネルをライフスタイル別に分類するための質問項目　約80問

因子分析（ケース1）



因子分析（ケース1主因子解・バリマックス回転）



クラスター分析(ケース1)

最終クラスタ中心 K-means

	クラスタ1	クラスタ2	クラスタ3	クラスタ4	クラスタ5
第１因子	-0.01	0.27	0.16	1.41	0.11
第２因子	0.18	1.32	-0.07	0.03	-0.05
第３因子	-0.35	0.11	-0.07	1.24	0.20
第４因子	0.16	0.23	-0.60	1.57	1.62
第５因子	-0.99	0.26	1.55	-0.04	1.41
第６因子	0.11	-0.38	0.07	1.54	0.30
第７因子	-0.09	-0.53	1.63	1.40	-0.40
第８因子	0.03	-0.08	0.25	1.61	0.23
第９因子	0.03	0.09	0.24	1.34	0.17
第１０因子	0.16	0.28	-0.08	1.65	-0.01
第１１因子	0.22	0.23	0.15	0.22	1.51
第１２因子	0.08	0.00	-0.04	0.11	0.15
第１３因子	0.17	0.24	1.61	0.03	0.25
第１４因子	-0.07	-0.06	0.23	0.19	0.28
第１５因子	-0.05	0.14	-0.10	-0.03	0.36
第１６因子	0.00	0.02	0.16	0.21	0.35
第１７因子	0.30	1.97	0.32	1.41	-0.66

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

クラスター分析(ケース1)

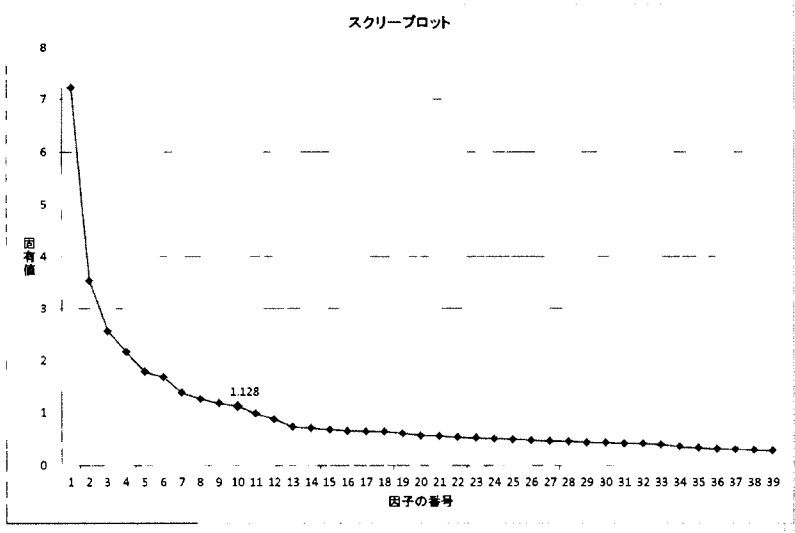
分散分析表

	クラスタ		誤差		F 値	有意確率
	平均平方	自由度	平均平方	自由度		
第１因子	71.44	4	0.95	5,624	75.20	0.00
第２因子	41.21	4	0.97	5,624	42.42	0.00
第３因子	69.04	4	0.95	5,624	72.55	0.00
第４因子	306.32	4	0.78	5,624	391.29	0.00
第５因子	483.13	4	0.66	5,624	735.25	0.00
第６因子	137.18	4	0.90	5,624	151.89	0.00
第７因子	277.16	4	0.80	5,624	344.91	0.00
第８因子	124.01	4	0.91	5,624	135.90	0.00
第９因子	54.40	4	0.96	5,624	56.55	0.00
第１０因子	133.42	4	0.91	5,624	147.30	0.00
第１１因子	115.74	4	0.92	5,624	126.03	0.00
第１２因子	11.65	4	0.99	5,624	11.74	0.00
第１３因子	153.01	4	0.89	5,624	171.55	0.00
第１４因子	46.00	4	0.97	5,624	47.52	0.00
第１５因子	42.23	4	0.97	5,624	43.51	0.00
第１６因子	50.35	4	0.96	5,624	52.19	0.00
第１７因子	476.16	4	0.66	5,624	719.22	0.00

各クラスタのケース数

クラスタ1	1,346
クラスタ2	1,116
クラスタ3	1,183
クラスタ4	953
クラスタ5	1,031
有効	5,629
欠損値	0

因子分析（ケース2主因子解・バリマックス回転）



因子分析（ケース2主因子解・バリマックス回転）

回転後の因子行列a	因子1	因子2	因子3	因子4	因子5	因子6	因子7	因子8	因子9	因子10	共通性
[redacted]	**0.79**	0.10	0.14	0.17	0.07	0.00	0.04	-0.12	-0.01	0.11	0.72
	0.78	0.12	0.16	0.14	0.12	-0.02	0.08	0.03	-0.01	0.05	0.70
	0.64	0.12	0.08	0.11	0.12	0.00	0.15	-0.01	-0.01	0.08	0.49
	0.62	0.13	0.22	0.22	0.13	0.02	0.01	-0.10	0.01	0.06	0.53
	0.53	0.07	0.27	0.20	0.08	-0.01	0.12	0.08	0.03	0.13	0.44
	0.09	**0.79**	0.15	0.09	0.03	0.01	0.00	0.10	-0.01	0.07	0.68
	0.07	**0.72**	0.15	0.12	-0.02	-0.10	0.03	0.07	0.03	-0.11	0.59
	0.10	**0.64**	0.15	0.05	-0.03	0.04	-0.07	0.12	-0.01	0.04	0.47
	0.08	**0.53**	0.17	0.03	0.04	0.00	-0.04	0.01	-0.07	0.04	0.45
	0.08	**0.56**	0.10	0.10	0.01	0.03	0.00	-0.01	0.02	-0.08	0.37
	0.05	**0.51**	0.03	0.10	0.05	-0.17	0.03	0.09	0.03	-0.10	0.33
	0.22	0.17	**0.67**	0.08	-0.01	0.00	-0.02	0.14	0.00	-0.02	0.55
	0.07	0.19	**0.57**	0.06	0.01	-0.02	-0.04	0.33	0.02	-0.06	0.48
	0.24	0.22	**0.56**	0.13	0.00	-0.02	-0.10	0.15	-0.05	-0.06	0.47
	0.19	0.11	**0.54**	0.12	0.07	0.03	0.10	-0.02	0.01	0.14	0.39
	0.10	0.11	**0.52**	0.17	0.00	0.03	-0.08	0.13	0.05	0.02	0.35
	0.00	0.10	**0.46**	0.12	-0.02	0.05	0.24	-0.14	0.11	0.08	0.33
	0.07	0.18	0.38	0.13	-0.06	-0.05	0.18	-0.36	0.12	-0.02	0.38
	0.25	0.12	0.19	**0.70**	-0.06	-0.03	-0.01	-0.04	0.01	0.01	0.62
	0.13	0.12	0.14	**0.69**	-0.05	0.01	0.12	0.19	0.00	0.06	0.58
	0.19	0.19	0.28	**0.65**	-0.07	0.04	0.07	0.03	0.08	0.04	0.60
	0.27	0.15	0.11	**0.64**	0.04	0.02	0.14	-0.02	0.03	0.09	0.55
	0.16	0.08	0.07	0.00	**0.66**	0.04	0.07	0.01	0.03	0.03	0.48
	0.16	0.05	0.04	-0.02	**0.59**	0.06	0.04	-0.01	0.00	0.21	0.43
	-0.06	-0.12	-0.04	-0.09	**0.56**	0.07	0.07	0.12	0.11	-0.11	0.40
	0.16	0.07	-0.04	0.01	**0.54**	0.01	0.10	0.08	0.04	0.12	0.36
	-0.01	-0.04	0.04	0.02	0.07	**0.74**	0.07	0.07	0.10	0.05	0.58
	0.00	-0.09	-0.02	0.01	0.06	**0.72**	0.02	-0.05	0.05	0.00	0.53
	0.01	0.00	0.02	-0.01	0.03	**0.53**	0.18	-0.03	0.02	0.05	0.32
	0.12	-0.09	-0.05	0.06	0.16	0.20	**0.74**	0.24	0.02	0.05	0.70
	0.17	-0.03	0.13	0.13	0.14	0.15	**0.69**	0.19	0.05	0.04	0.58
	0.38	0.02	-0.04	0.20	0.16	0.13	**0.44**	-0.06	0.03	0.10	0.44
	-0.07	0.10	0.02	0.05	0.03	0.02	0.07	**0.49**	0.18	-0.01	0.30
	0.00	0.14	0.24	0.05	0.06	0.00	0.13	**0.46**	0.08	0.03	0.32
	-0.03	0.13	0.16	0.04	0.09	-0.08	0.14	**0.44**	0.08	0.02	0.28
	0.00	0.02	0.07	0.06	0.07	0.10	0.04	0.14	**0.72**	-0.02	0.56
	0.01	-0.05	0.04	0.01	0.09	0.07	0.02	0.12	**0.71**	0.03	0.53
	0.34	0.01	0.10	0.12	0.20	0.04	0.01	-0.06	0.01	**0.60**	0.55
	0.10	-0.14	0.01	0.06	0.07	0.08	0.08	0.07	0.00	**0.59**	0.41
固有値	3.11	2.99	2.57	2.21	1.63	1.52	1.49	1.24	1.13	0.93	18.83
累積寄与率(%)	7.97	15.65	22.24	27.90	32.08	35.97	39.80	42.98	45.88	48.28	

因子抽出法: 主因子法　回転法: Kaiser の正規化を伴うバリマックス法
a. 23 回の反復で回転が収束しました。

クラスター分析（ケース2）

最終クラスタ中心 K-means

	クラスタ1	クラスタ2	クラスタ3	クラスタ4	クラスタ5
因子1	-0.46	-0.03	-0.56	-0.15	1.20
因子2	0.24	-0.32	-0.01	-0.05	0.13
因子3	0.74	-1.01	0.16	-0.22	0.26
因子4	0.03	-0.03	-0.13	-0.28	0.39
因子5	-0.26	-0.01	-0.03	0.01	0.29
因子6	-0.51	-0.28	-0.10	1.09	-0.08
因子7	-0.65	-0.12	1.08	-0.66	0.20
因子8	-0.35	-0.66	0.30	0.64	0.11
因子9	0.07	-0.27	0.07	0.03	0.10
因子10	-0.05	0.04	0.01	-0.14	0.14

クラスター分析（ケース2）

分散分析表

	クラスタ		誤差		F 値	有意確率
	平均平方	自由度	平均平方	自由度		
因子1	575.93	4	0.59	5,624	974.35	0.00
因子2	51.32	4	0.96	5,624	53.23	0.00
因子3	480.25	4	0.66	5,624	728.60	0.00
因子4	69.12	4	0.95	5,624	72.64	0.00
因子5	44.71	4	0.97	5,624	46.15	0.00
因子6	398.84	4	0.72	5,624	556.23	0.00
因子7	598.35	4	0.58	5,624	1040.34	0.00
因子8	291.26	4	0.79	5,624	367.03	0.00
因子9	26.72	4	0.98	5,624	27.21	0.00
因子10	12.03	4	0.99	5,624	12.13	0.00

各クラスタのケース数	
クラスタ1	1,139
クラスタ2	1,115
クラスタ3	1,212
クラスタ4	1,007
クラスタ5	1,156
有効	5,629
欠損値	0

因子得点の計算_因子得点平均と分散

		A-R factor score		REGR factor score		BART factor score	
	度数	平均値	分散	平均値	分散	平均値	分散
因子1	5,629	0.00	1.00	0.00	0.82	0.00	1.27
因子2	5,629	0.00	1.00	0.00	0.83	0.00	1.22
因子3	5,629	0.00	1.00	0.00	0.74	0.00	1.38
因子4	5,629	0.00	1.00	0.00	0.76	0.00	1.34
因子5	5,629	0.00	1.00	0.00	0.69	0.00	1.50
因子6	5,629	0.00	1.00	0.00	0.74	0.00	1.38
因子7	5,629	0.00	1.00	0.00	0.73	0.00	1.42
因子8	5,629	0.00	1.00	0.00	0.62	0.00	1.67
因子9	5,629	0.00	1.00	0.00	0.68	0.00	1.48
因子10	5,629	0.00	1.00	0.00	0.59	0.00	1.73

17

因子得点の計算_因子間の相関

アンダーソン・ルービン法

Pearson の相関係数	因子1	因子2	因子3	因子4	因子5	因子6	因子7	因子8	因子9	因子10
因子1	1	.000	.000	.000	.000	.000	.000	.000	.000	.000
因子2	.000	1	.000	.000	.000	.000	.000	.000	.000	.000
因子3	.000	.000	1	.000	.000	.000	.000	.000	.000	.000
因子4	.000	.000	.000	1	.000	.000	.000	.000	.000	.000
因子5	.000	.000	.000	.000	1	.000	.000	.000	.000	.000
因子6	.000	.000	.000	.000	.000	1	.000	.000	.000	.000
因子7	.000	.000	.000	.000	.000	.000	1	.000	.000	.000
因子8	.000	.000	.000	.000	.000	.000	.000	1	.000	.000
因子9	.000	.000	.000	.000	.000	.000	.000	.000	1	.000
因子10	.000	.000	.000	.000	.000	.000	.000	.000	.000	1

回帰法

Pearson の相関係数	因子1	因子2	因子3	因子4	因子5	因子6	因子7	因子8	因子9	因子10
因子1	1	.023	.073**	.078**	.080**	-.014	.048**	-.068**	-.017	.082**
因子2	.023	1	.074**	.037**	.010	-.023	-.029*	.045**	-.014	-.031*
因子3	.073**	.074**	1	.068**	-.018	.008	-.024	.081**	.026*	.003
因子4	.078**	.037**	.068**	1	-.064**	.001	.053**	.006	.026	.037**
因子5	.080**	.010	-.018	-.064**	1	.033*	.067**	.053**	.047**	.084**
因子6	-.014	-.023	.008	.001	.033*	1	.076**	-.005	.048**	.040**
因子7	.046**	-.029*	-.024	.053**	.067**	.076**	1	.108**	.004	.009
因子8	-.068**	.045**	.081**	.006	.053**	-.005	.108**	1	.075**	-.017
因子9	-.017	-.014	.026*	.026	.047**	.048**	.004	.075**	1	-.017
因子10	.082**	-.031*	.003	.037**	.084**	.040**	.009	-.017	-.017	1

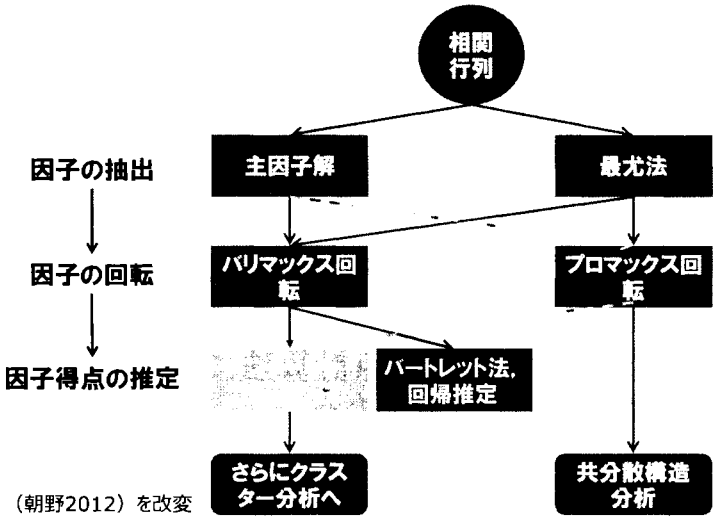
バートレット法

Pearson の相関係数	因子1	因子2	因子3	因子4	因子5	因子6	因子7	因子8	因子9	因子10
因子1	1	-.021	-.078**	-.074**	-.082**	.024	-.051**	.081**	.016	-.073**
因子2	-.002	1	-.066**	-.035**	-.016	.018	.034*	-.044**	.020	.035**
因子3	-.078**	-.066**	1	-.058**	.025	-.014	.038**	-.087**	-.021	.000
因子4	-.074**	-.035**	-.058**	1	.077**	.002	-.056**	-.001	-.030*	-.038**
因子5	-.082**	-.016	.025	.077**	1	-.026	-.059**	-.052**	-.048**	-.081**
因子6	.002	.018	.014	.002	-.026	1	-.077**	.019	-.048**	-.040**
因子7	-.051**	.034*	.038**	-.056**	-.059**	-.077**	1	-.113**	.012	.003
因子8	.081**	-.044**	-.087**	-.001	-.052**	.019	.113**	1	-.070**	.013
因子9	.002	.020	-.021	-.030*	-.048**	-.048**	.012	-.070**	1	.022
因子10	-.073**	.035**	.000	-.038**	-.081**	-.040**	.005	.013	.022	1

**. 相関係数は 1% 水準で有意（両側）です。
*. 相関係数は 5% 水準で有意（両側）です。

18

因子分析のフロー



（朝野2012）を改変

3. 因子の回転

- バリマックス回転
  - 相関がゼロという直行条件を付けて回転を行う直交回転
  - 因子負荷量が-1～+1の間までの値を取り、理解しやすい
  - 縦横90度に回転するので、グラフの表示が簡単
  - マーケティング等で使用されやすい

- プロマックス回転
  - 因子間の相関を許容する斜交回転
  - 因子の解釈度が高くなる
  - 共分散構造分析のために探索的に因子分析をする際は良い
  - 心理学などアカデミックな分野で使用されやすい

（朝野2012）

因子分析の手順

データ：ネットリサーチ・パネル属性調査

- 因子抽出：主因子法
- 因子数：固有値>1
- 回転：バリマックス回転とプロマックス回転(k=4)
- 因子得点：回帰法、Bartlett法、Anderson-Rubin法
- それぞれの回転と因子得点計算法によって、因子得点毎の相関行列を求める

バリマックス回転後の因子例

回転後の因子行列a

	1	2	3	4	5	6	7	8	9	10	共通性
	.612	.087	-.113	.167	.061	.018	.005	-.026	.067	.117	0.73
	.777	.120	.167	.143	.123	.090	-.022	-.007	.038	.095	0.69
	.646	.120	.068	.106	.126	.147	-.009	-.014	.054	.075	0.49
	.836	.122	.183	.216	.123	-.008	.017	-.004	.096	.064	0.53
	.612	.075	.274	.199	.075	.149	-.006	.032	.075	.136	0.44
	.080	.801	.158	.089	.032	.011	.010	-.002	.042	.073	0.69
	.066	.716	.140	.128	-.009	.016	-.103	.029	.080	-.118	0.59
	.085	.645	.174	.048	-.027	-.040	.056	-.007	.019	.044	0.46
	.084	.696	.159	.033	.038	-.038	-.002	-.069	.056	.050	0.45
	.084	.674	.093	.099	.014	-.012	.013	.014	.057	-.075	0.36
	.048	.614	.044	.109	.068	.021	-.169	.032	.020	-.108	0.33
	.194	.174	.656	.083	.007	.025	-.016	.025	.150	.006	0.57
	.024	.211	.692	.074	.021	.053	-.023	.063	.035	.034	0.48
	.204	.226	.623	.136	-.004	-.033	-.030	-.021	.038	-.029	0.50
	.071	.116	.566	.176	.000	-.024	.019	.074	.076	.045	0.37
	.196	.105	.454	.125	.069	.070	.027	.010	.273	.147	0.38
	.258	.116	.167	.704	.066	.016	-.022	.002	.071	.017	0.61
	.109	.130	.164	.690	.038	.153	.003	.019	.035	.059	0.55
	.194	.187	.248	.655	.073	.072	.029	.078	.125	.050	0.60
	.279	.141	.066	.643	.038	.115	.018	.020	.122	.089	0.55
	.164	.074	.056	.003	.653	.061	.042	.019	.046	.036	0.48
	.168	.052	.028	.025	.583	.048	.063	.002	.012	.222	0.43
	.070	.115	.006	.087	.596	.102	.053	.131	.069	.107	0.40
	.152	.072	.031	.011	.547	.116	.002	.050	.035	.119	0.36
	.083	.064	.023	.060	.159	.843	.117	.057	.013	.061	0.77
	.142	.008	.123	.134	.140	.712	.094	.078	.062	.053	0.60
	.388	.021	.078	.203	.157	.420	.072	.025	.076	.102	0.42
	.012	.090	.039	.006	.058	.065	.783	.042	.010	.014	0.66
	.010	.036	.021	.014	.075	.128	.792	.107	.033	.061	0.54
	.004	.035	.041	.012	.090	.042	.049	.746	.021	.034	0.58
	.014	.032	.060	.059	.076	.071	.094	.712	.035	.017	0.54
	.046	.061	.233	.102	.002	.101	.067	.033	.649	.056	0.51
	.157	.137	.107	.124	.053	.017	.029	.028	.586	.052	0.42
	.341	.012	.071	.121	.190	.008	.023	.007	.042	.617	0.56
	.089	.128	.008	.059	.072	.108	.055	.012	.013	.584	0.39
因子寄与	3.113	2.943	2.270	2.192	1.610	1.575	1.173	1.139	.972	.952	17.94
累積寄与率(%)	6.994	17.302	23.788	30.050	34.650	39.192	42.513	45.768	48.544	51.265	

因子抽出法： 主因子法 回転法： Kaiser の正規化をともなうバリマックス法

a. 7 回の反復で回転が収束しました。

バリマックス回転での因子得点の相関行列

回帰法

因子	1	2	3	4	5	6	7	8	9	10
1	1	.017	.050**	.079**	.070**	.017	-.012	-.029*	.032*	.085**
2	.017	1	.078**	.038**	.017	-.018	-.023	-.009	.023	-.032*
3	.050**	.078**	1	.055**	-.006	-.006	-.004	.034**	.109**	.004
4	.079**	.038**	.055**	1	-.063**	.043**	.002	.024	.052**	.036**
5	.070**	.017	-.006	-.063**	1	.076**	.031*	.055**	-.026*	.086**
6	.017	-.018	-.006	.043**	.076**	1	.052**	.025	.014	.013
7	-.012	-.023	-.004	.002	.031*	.052**	1	.045**	.026*	.039**
8	-.029*	-.009	.034**	.024	.055**	.025	.045**	1	.018	-.013
9	.032*	.023	.109**	.052**	-.026*	.014	.026*	.018	1	.010
10	.085**	-.032*	.004	.036**	.086**	.013	.039**	-.013	.010	1

Bartlett法

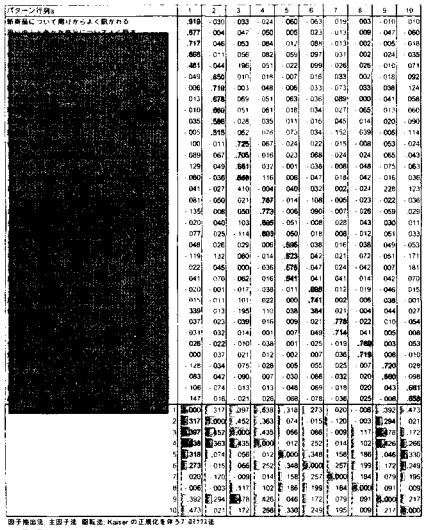
因子	1	2	3	4	5	6	7	8	9	10
1	1	-.011	-.043**	-.077**	-.071**	-.009	.017	.035**	-.025*	-.077**
2	-.011	1	-.074**	-.036**	-.025	.019	.021	.013	-.015	.036**
3	-.043**	-.074**	1	-.042**	.006	.009	.006	-.035**	-.103**	-.001
4	-.077**	-.036**	-.042**	1	.076**	-.047**	.000	-.028*	-.041**	-.036**
5	-.071**	-.025	.006	.076**	1	-.076**	-.024	-.059**	.028*	-.083**
6	-.009	.019	.009	-.047**	-.076**	1	-.048**	-.018	-.013	-.002
7	.017	.021	.006	.000	-.024	-.048**	1	-.042**	-.027*	-.037**
8	.035**	.013	-.035**	-.028*	-.059**	-.018	-.042**	1	-.015	.019
9	-.025*	-.015	-.103**	-.041**	.028*	-.013	-.027*	-.015	1	-.008
10	-.077**	.036**	-.001	-.036**	-.083**	-.002	-.037**	.019	-.008	1

Anderson-Rubin法

因子	1	2	3	4	5	6	7	8	9	10
1	1	.000	.000	.000	.000	.000	.000	.000	.000	.000
2	.000	1	.000	.000	.000	.000	.000	.000	.000	.000
3	.000	.000	1	.000	.000	.000	.000	.000	.000	.000
4	.000	.000	.000	1	.000	.000	.000	.000	.000	.000
5	.000	.000	.000	.000	1	.000	.000	.000	.000	.000
6	.000	.000	.000	.000	.000	1	.000	.000	.000	.000
7	.000	.000	.000	.000	.000	.000	1	.000	.000	.000
8	.000	.000	.000	.000	.000	.000	.000	1	.000	.000
9	.000	.000	.000	.000	.000	.000	.000	.000	1	.000
10	.000	.000	.000	.000	.000	.000	.000	.000	.000	1

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

プロマックス回転後の因子例



パターン行列a

	1	2	3	4	5	6	7	8	9	10
新商品について周りからよく聞かれる	.919	.030	.033	-.024	.060	-.063	.019	.003	-.010	.010
	.877	.004	.047	.050	.005	.023	-.013	.009	.047	.060
	.717	.046	.053	.084	.012	.084	-.013	.002	.005	.048
	.608	.011	.056	.082	.059	.097	.031	.002	.024	.035
	.481	.044	.196	.051	.022	.099	.026	.026	.010	.071
	.049	.850	.010	.016	.007	.016	.033	.002	.018	.092
	.006	.716	.001	.048	.006	.033	.073	.033	.038	.124
	.013	.678	.069	.051	.063	.036	.089	.000	.041	.058
	.010	.660	.051	.061	.018	.034	.027	.065	.019	.060
	.035	.586	.028	.035	.011	.016	.045	.014	.020	.090
	.005	.515	.062	.026	.073	.034	.152	.039	.005	.114
	.100	.011	.725	.067	.024	.022	.015	.058	.053	.024
	.089	.067	.705	.016	.023	.068	.024	.024	.065	.043
	.129	.049	.681	.032	.001	.038	.008	.048	.075	.063
	.060	.036	.669	.116	.006	.047	.018	.042	.016	.036
	.041	.027	.419	.004	.040	.032	.002	.024	.228	.123
	.061	.009	.021	.767	.014	.198	.005	.023	.022	.036
	.135	.008	.050	.773	.006	.060	.007	.026	.059	.029
	.020	.040	.103	.685	.051	.008	.028	.043	.030	.011
	.077	.025	.114	.669	.050	.018	.008	.012	.051	.033
	.048	.026	.029	.006	.886	.038	.016	.038	.049	.053
	.119	.132	.060	.074	.823	.042	.021	.072	.051	.171
	.022	.045	.000	.036	.676	.047	.024	.042	.007	.181
	.041	.070	.052	.016	.641	.041	.041	.014	.042	.070
	.020	.001	.017	.038	.011	.058	.012	.019	.046	.015
	.015	.011	.161	.022	.000	.741	.002	.006	.038	.001
	.339	.013	.165	.110	.038	.384	.021	.004	.044	.027
	.031	.023	.039	.016	.009	.021	.778	.022	.010	.054
	.031	.032	.014	.001	.007	.049	.714	.041	.005	.008
	.028	.022	.010	.038	.001	.025	.013	.760	.003	.053
	.000	.031	.021	.012	.002	.007	.036	.719	.006	.015
	.148	.034	.075	.028	.005	.055	.025	.007	.720	.029
	.053	.042	.090	.007	.030	.066	.032	.029	.680	.098
	.106	.074	.013	.013	.048	.069	.018	.020	.043	.681
	.147	.018	.021	.025	.008	.078	.036	.025	.008	.658

	1	2	3	4	5	6	7	8	9	10
1	1.000	.517	.297	.438	.318	.273	.020	.008	.392	.472
2	.517	1.000	.452	.363	.074	.015	.120	.003	.294	.021
3	.297	.452	1.000	.435	.056	.066	.009	.117	.078	.172
4	.438	.363	.435	1.000	.012	.252	.014	.102	.226	.266
5	.318	.074	.056	.012	1.000	.348	.158	.186	.046	.330
6	.273	.015	.066	.252	.348	1.000	.057	.199	.172	.249
7	.020	.120	.009	.014	.158	.257	1.000	.184	.079	.195
8	.008	.003	.117	.102	.186	.199	.184	1.000	.091	.009
9	.392	.294	.078	.226	.046	.172	.079	.091	1.000	.217
10	.472	.021	.172	.266	.330	.249	.195	.009	.217	1.000

因子抽出法: 主因子法 回転法: Kaiser の正規化を伴うプロマックス法
a. 7 個の反復で回転が収束しました。

プロマックス回転での因子得点の相関行列

回帰法

因子	1	2	3	4	5	6	7	8	9	10
1	1	.352**	.458**	.600**	.371**	.307**	.028*	.001	.477**	.571**
2	.352**	1	.525**	.414**	.085**	-.015	-.143**	.001	.370**	.035**
3	.458**	.525**	1	.512**	.075**	.083**	-.016	.143**	.598**	.216**
4	.600**	.414**	.512**	1	.028*	.286**	.021	.120**	.532**	.338**
5	.371**	.085**	.075**	.028*	1	.417**	.208**	.232**	.071**	.428**
6	.307**	-.015	.083**	.286**	.417**	1	.317**	.241**	.211**	.319**
7	.028*	-.143**	-.016	.021	.208**	.317**	1	.240**	.097**	.240**
8	.001	.001	.143**	.120**	.232**	.241**	.240**	1	.125**	.032*
9	.477**	.370**	.598**	.532**	.071**	.211**	.097**	.125**	1	.286**
10	.571**	.035**	.216**	.338**	.428**	.319**	.240**	.032*	.286**	1

Bartlett法

因子	1	2	3	4	5	6	7	8	9	10
1	1	.281**	.332**	.469**	.259**	.238**	.015	-.007	.310**	.369**
2	.281**	1	.372**	.310**	.059**	-.013	-.097**	-.004	.223**	.016
3	.332**	.372**	1	.354**	.041**	.051**	-.002	.088**	.347**	.134**
4	.469**	.310**	.354**	1	.012	.216**	.009	.083**	.320**	.195**
5	.259**	.059**	.041**	.012	1	.275**	.111**	.136**	.034*	.226**
6	.238**	-.013	.051**	.216**	.275**	1	.197**	.160**	.135**	.184**
7	.015	-.097**	-.002	.009	.111**	.197**	1	.130**	.056**	.149**
8	-.007	-.004	.088**	.083**	.136**	.160**	.130**	1	.059**	-.006
9	.310**	.223**	.347**	.320**	.034*	.135**	.056**	.059**	1	.158**
10	.369**	.016	.134**	.195**	.226**	.184**	.149**	-.006	.158**	1

Anderson-Rubin法

因子	1	2	3	4	5	6	7	8	9	10
1	1	.000	.000	.000	.000	.000	.000	.000	.000	.000
2	.000	1	.000	.000	.000	.000	.000	.000	.000	.000
3	.000	.000	1	.000	.000	.000	.000	.000	.000	.000
4	.000	.000	.000	1	.000	.000	.000	.000	.000	.000
5	.000	.000	.000	.000	1	.000	.000	.000	.000	.000
6	.000	.000	.000	.000	.000	1	.000	.000	.000	.000
7	.000	.000	.000	.000	.000	.000	1	.000	.000	.000
8	.000	.000	.000	.000	.000	.000	.000	1	.000	.000
9	.000	.000	.000	.000	.000	.000	.000	.000	1	.000
10	.000	.000	.000	.000	.000	.000	.000	.000	.000	1

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

クラスター分析 K-means

最終クラスター中心

バリマックス回転

因子	クラスタ 1	2	3	4	5
1	.69411	-.[illegible]3638	-.04297	.04087	-.[illegible]2414
2	.25110	.3[illegible]061	-.[illegible]0135	.37[illegible]72	.12348
3	.03970	.3[illegible]528	-.[illegible]2390	-.2[illegible]526	.02829
4	.65[illegible]29	.4[illegible]061	.01453	-.[illegible]583	-.1[illegible]624
5	.67[illegible]16	-.[illegible]8758	-.03271	.17462	.03126
6	.07336	.1[illegible]031	-.10741	-.06064	-.1[illegible]011
7	-.00812	-.[illegible]5754	-.[illegible]2460	-.[illegible]766	1.30[illegible]16
8	-.01105	-.[illegible]9406	-.01748	.17524	.13389
9	.28026	.1[illegible]663	.00036	-.2[illegible]644	-.2[illegible]616
10	.32067	.1[illegible]589	-.03222	-.2[illegible]227	-.1[illegible]891

プロマックス回転

因子	クラスタ 1	2	3	4	5
1	.77[illegible]93	-.[illegible]8753	-.00329	.01829	-.[illegible]2229
2	.23[illegible]31	.3[illegible]248	-.[illegible]9271	.38[illegible]12	.13464
3	-.01161	.3[illegible]148	-.2[illegible]848	-.2[illegible]026	.04370
4	.62[illegible]21	.4[illegible]846	-.01795	-.[illegible]634	-.1[illegible]206
5	.63[illegible]16	.[illegible]5667	-.00921	.15424	.04281
6	.12167	.10118	-.08443	-.08109	-.01891
7	-.00079	-.[illegible]8237	-.2[illegible]793	-.[illegible]671	1.30[illegible]63
8	-.00232	-.[illegible]1693	-.01422	.18793	.13076
9	.23[illegible]70	.1[illegible]865	.02178	-.2[illegible]967	-.2[illegible]211
10	.30[illegible]63	.10483	-.01177	-.2[illegible]891	-.1[illegible]610

因子得点はそれぞれAnderson-Rubin法

クラスター分析 K-means

各クラスター間の距離

バリマックス回転

クラスタ	1	2	3	4	5
1		2.015	2.280	1.963	2.130
2	2.015		2.346	1.935	2.166
3	2.280	2.346		2.203	2.365
4	1.963	1.935	2.203		2.050
5	2.130	2.166	2.365	2.050	

プロマックス回転

クラスタ	1	2	3	4	5
1		2.015	2.280	1.963	2.130
2	2.015		2.346	1.935	2.166
3	2.280	2.346		2.203	2.365
4	1.963	1.935	2.203		2.050
5	2.130	2.166	2.365	2.050	

各クラスターのケース数

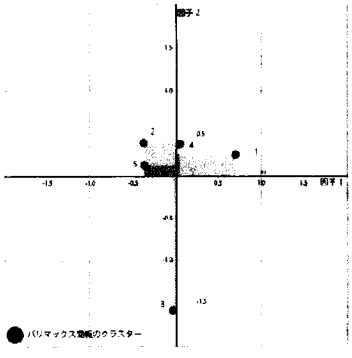
クラスタ	1	1198.000
	2	1175.000
	3	873.000
	4	1329.000
	5	1054.000
有効		5629.000
欠損値		0.000

クラスタ	1	1198.000
	2	1175.000
	3	873.000
	4	1329.000
	5	1054.000
有効		5629.000
欠損値		0.000

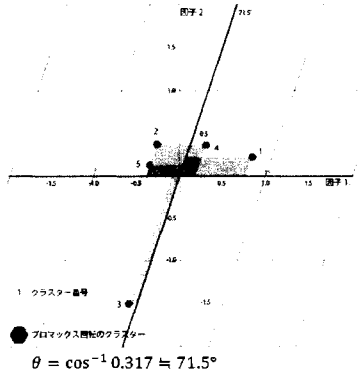
		プロマックスのクラスター					合計
		1	2	3	4	5	
バリマックス	1	1198	0	0	0	0	1198
のクラスター	2	0	1175	0	0	0	1175
	3	0	0	873	0	0	873
	4	0	0	0	1329	0	1329
	5	0	0	0	0	1054	1054
合計		1198	1175	873	1329	1054	5629

回転ごとのプロット図

バリマックス回転

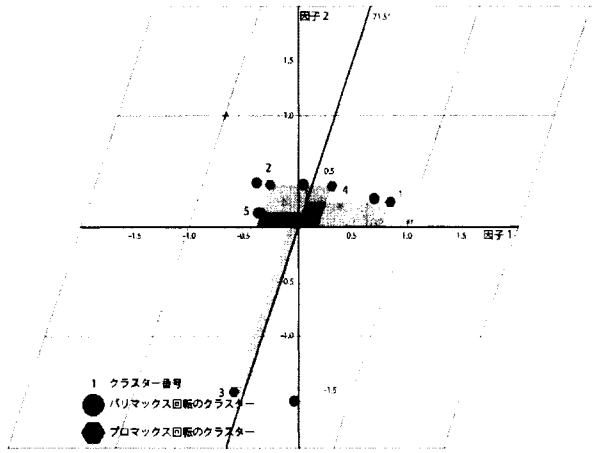


プロマックス回転



$\theta = \cos^{-1} 0.317 \fallingdotseq 71.5°$

重ね合わせ



医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

G4：コーホート分析

田村玄
株式会社 ビデオリサーチ

G4: Cohort analysis

Gen Tamura
Video Research Ltd.

時代のトレンドをとらえる

コーホート分析概要
- トレンドを年齢・時代・コーホート(世代)に分解
- トレンドの構造が明らかになる
- 長期間、年齢別に観測されたデータが必要

コーホート分析アルゴリズム
- ベイズ型コウホートモデル、簡易推計、重回帰
- SAS/IMLを用いた簡易推計はベイズ型コウホートモデルと著しく異なるわけではない
- 重回帰は、多重共線性をどう回避させるかが課題

コーホート分析による将来予測
- コーホート分析はフィッティング重視
- 単純な自己回帰の方が誤差が少ないケースもある

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

「マーケティング分野におけるベイズ統計の活用事例に関する一考察」

中見真也 学習院大学 経営学研究科 博士後期課程
松本和宏 (株)富士通研究所 ナレッジプラットフォーム研究部

A study of the application case of Bayesian statistics in the marketing field.

Shinya Nakami, Gakushuin University
Kazuhiro Matsumoto, FUJITSU LABORATORIES LTD.
KNOWLEDGE PLATFORMS LAB.

AGENDA

- 1. ベイズ統計の概要
  - 1-1. ベイズ統計の基本的な考え方
  - 1-2. 頻度主義とベイズ統計の違い
  - 1-3. ベイズ統計のメリット・デメリット
- 2. ベイズ統計を理解する上で重要なキーワード
  - 2-1. マルコフ連鎖モンテカルロ法(MCMC法)
  - 2-2. 複雑な統計モデルに対応する階層ベイズ法
- 3. 分析事例～階層ベイズと線形回帰の違い
- 4. 課題と展望～実務への応用について

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

ベイズ統計の基本的な考え方

■そもそもベイズって？

18世紀後半スコットランドの長老派教会の牧師トーマス・ベイズが考案した「ベイズの定理」が由来.
(つまり、アマチュアの数学者)

■ベイズの定理とは？

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

事象Aに関するある結果(データ)が得られたとすると、それを反映し、尤度P(A/B)の乗算によって、事象Bの確率は事前確率から事後確率へと更新される。

事前確率 → 新データ → 新たな確率

得たデータが前の確率に作用して新たな確率が算出される(ベイズ更新)
※主観確率と言われる所以.(つまり経験を取り込むことができる)

ベイズ統計を使う理由(メリット/デメリット)

＜メリット＞
①柔軟かつ自然なモデリングが実現できる
②小サンプルでも，妥当なパラメータが推定できる
※標本数が少ないことに起因する不適解を回避できる場合がある
③誤差分散が負になるといった、計算上のエラーを避けられる

＜デメリット＞
①計算が結構大変
②事前分布の設定により客観性を損なう恐れがある

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

課題と展望

●今日、実務上、ベイズ統計の考え方をマーケティング活動に利用している事例は正直少ないのが現状。
→まずは、ベイズ統計と頻度主義を併用してみて、どちらが有効かを試してみることが重要！！

●製造業、サービス業を中心に、膨大な顧客データを今後どうビジネスに活用すべきか、正にビックデータ時代のマーケティングを考える上で、ベイズ統計の考え方は、時代にフィットする可能性を秘めている。

●消費者視点で購買行動を見た際にも、現在スマホ、PCの普及により、リアルとネットを自由に行き来する「オムニチャネル」が今後普及していくことが予測される。その際に、個人の購買履歴をベースに、企業側は関連販売（リコメンデーション）を推進していく際にもベイズ統計は有効だと考えられる。

リスク管理

Domination理論によるリスク管理標準

中西　美紗
統計数理研究所リスク解析戦略研究センター

Domination theory on general graphs for risk management

Misa Nakanishi
Risk analysis research center, The institute of statistical mathematics

要旨

リスクマネジメントは、予期しない事象の影響を評価し、資質の効用を促進する。その標準モデルとして、組織体をグラフ表現し domination を一般化する。これは、set covering problem と関連し、本論では SAS programming による計算を記述する。

キーワード：リスクマネジメント、Domination theory、graphical representation、SQL procedure

---

序論

リスクマネジメントは、予期しない事象の影響を評価し、資質の効用を促進する。その標準モデルとして、組織体をグラフ表現する。構成要素をノード（頂点）化し、その間の関係を辺にすることにより、グラフを定義する。グラフの dominating set は、頂点集合の部分集合であり、全ての頂点を cover する。信用リスク、情報ネットワーク、災害政策など、現実のシーンへの応用は多様だ。domination theory は set covering problem と関連し、本論ではこの問題への解を与えることを主旨として、SAS programming による計算を記述する。

位相空間を基礎として、set covering problem は、集合の部分集合からなる family のうち、集合全体の要素を cover する最小数の部分集合を求めようとする。これと同値な表現として、bipartite graph へ写像し、domination number を与えることが、中心的な命題である。よって、その間の帰結は NP-complete として同等な任意のグラフの dominating set を導く。

Domination model におけるリスク管理

リスクマネジメントにおいて、組織体の要素を覆う集合を扱う標準モデルを構成し全体を俯瞰することが求められる。グラフ表現による Domination 性の付与は普遍的である。

"the effect of uncertainty on objectives" と記されたリスクマネジメントは、ISO 31000 により 3 つの場面 - identification、assessment、prioritization を定めている [1]。全てのリスクの存在を評価するのは難しく、資質を均衡を保ちながら配置するのに多くの失敗を伴う。構成要素のつながりの上で、それらを網羅する集約点をそれぞれに形成し、全てのリスクに適応することが基本的である。経費を最小とし、さらにリスクの負の効果を最小にするのは、opportunity cost の考え方に通じている。

適応する現実のシーンとして、次に例を挙げる [2]。
project management security engineering industrial processes financial portfolios actuarial assessments public health and safety

最近には、Intangible risk management と呼ぶ、必ず起こる事象が identification の欠如により組織に見落とされるリスクの型が提唱されている。Domination 性の付与は、これまで見落とされてきたそれらの資質を有効化する。

グラフ表現

一般にグラフは頂点集合と辺集合から定義する。辺集合は頂点の二項関係からなる [3]。

$$G = (V, E)$$

頂点集合 V と辺集合 E を SAS dataset として保持すると次のように表される。

vertex
a
b
c
d
e
f

data vertex

var1	var2
a	b
a	e
b	c
b	d
b	e
c	e
d	a
d	f
e	f
f	a
f	b

data edge

図 1　グラフの頂点集合と辺集合

0. Domination model

一般にグラフにおける dominating set とは、その補集合の頂点を全て cover する頂点集合である。つまり、グラフ G に対して、dominating set X は V の部分集合で V – X の頂点はそれぞれ X の頂点と隣接している。
図 1 のグラフを例に、ひとつの dominating set を図 2 に示す。

vertex
a
b
c
d
e
f

data vertex

var1	var2
a	b
a	e
b	c
b	d
b	e
c	e
d	a
d	f
e	f
f	a
f	b

data edge

図 2　dominating set X = {a, c}

グラフの dominating set の頂点とそれぞれの隣接頂点は元の頂点集合を形成する。それにより、dominating set であるかを確かめる。図 2 において、SQL procedure を用い、次の(1)と(2)から成る data は(3)の data と等しい。

(1) select var2 from edge where var1 = 'a' or var1 = 'c';

(2) select var1 from edge where var2 = 'a' or var2 = 'c';

(3) select vertex from vertex where vertex ^= 'a' and vertex ^= 'c';

dominating set のいくつかの特徴的な例を次に示す。

0-1.　Connected domination

グラフが connected であるとき、任意の 2 頂点は辺でつながった頂点を経由しながら互いに移りあう。dominating set に対しても同様に、connected であるという。connected dominating set は、例えば mobile ad hoc network の routing の考察に利用され、communication の基幹にある [4]。

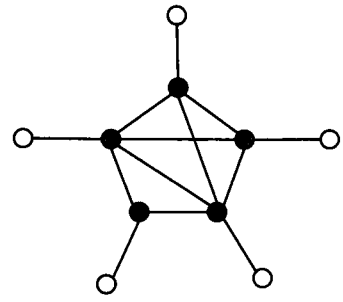


図 3　グラフの connected dominating set

図 1 のように、頂点集合と辺集合の dataset を作成し、dominating set を X とおく。

(4) select var1, var2 from edge where var1 in X and var2 in X;

これは、X の間の辺を表す。connected であるならば、X の頂点の隣接点の集合 Y に対して、(4)の table に繰り返し(5)、(6)を適用し、X の全ての頂点を抽出する。

(5) select var2 from (4) where var1 in Y;

(6) select var1 from (4) where var2 in Y;

0-2.　Independent domination

dominating set が independent であるとき、その任意の 2 頂点は辺をもたない。これは、0-1 の(4)の施行による observation 数が 0 であることと同等である。

0-3. k - domination

自然数 k に対して、X が k-dominating set ならば、V - X の頂点がそれぞれ X の少なくとも k 頂点と隣接する。

図 1 のように、辺集合の dataset を作成し、dominating set を X とおく。X の補集合の頂点 v に対して、

(7) select var2 from edge where var1 = 'v' and var2 in X;

(8) select var1 from edge where var2 = 'v' and var1 in X;

を施行すると、(7)と(8)の observation の総数が k 以上である。

構成

グラフの domination model を構成する。図 1 のような、頂点集合と辺集合の dataset に対して、SQL procedure を施行する。

(1) select var2 from edge where var1 = 'a';

(2) select var1 from edge where var2 = 'a';

(3) delete from edge where var1 = 'a' or var2 = 'a' or (var1 in (1) and var2 in (1)) or (var1 in (2) and var2 in (2)) or (var1 in (1) and var2 in (2)) or (var1 in (2) and var2 in (1));

vertex a について、(3)の table を決め、それに対して繰り返し 0 になるまで同様に vertex を取り出す。その頂点集合は dominating set である。

var1	var2
a	b
a	e
b	c
b	d
b	e
c	e
d	a
d	f
e	f
f	a
f	b

var1	var2
b	c
c	e

var1	var2
c	e

図 4　dominating set {a, b, c}の構成

A minimum dominating set

一般のグラフにおいて、頂点数が最小の dominating set を構成するのは、NP-hard であるとされる。decision problem のひとつで、次のような関連する同等の問題を示している [5]。

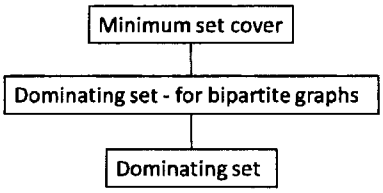


図 5　NP-complete problem の関連

グラフ G の minimum dominating set の位数を domination number といい、γ(G)と記す。

Minimum set cover と Dominating set

集合 S = {a_1, ..., a_m} の部分集合 C_1, ..., C_n が与えられ、それらのある和集合が S を覆うとき、それを満たす部分集合の最小数を求める。その最小数がある値未満であるか、Minimum set cover は提起する。これは、次のグラフ G0 において、domination number を導くことに帰着する [6]。

G0 = (V0, E0)

V0 = {v, w} ∪ {C_1, ..., C_n} ∪ {a_1, ..., a_m}

E0 = {(v, w)} ∪ {(v, C_i) | i = 1, ..., n} ∪ {(C_i, a_k) | a_k in C_i, i = 1, ..., n}

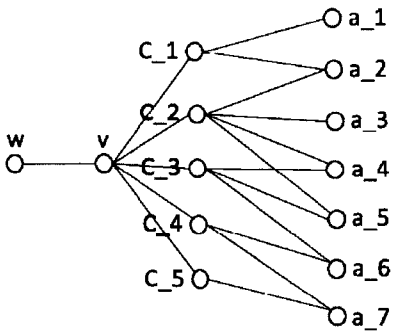


図 6　Minimum set cover の instance のグラフ表現

このグラフはひとつの bipartite graph を構成している。independent で互いに素な 2 つの頂点集合からなる。bipartite graph の independent domination number の解法がこれまでに示唆されている [7]。グラフ G の independent domination number は、independent dominating set の最小位数であり、i(G)と記す。

解法

Minimum set cover の instance のグラフ表現について、domination number を与える方法を概略する。グラフ G0 について、次の過程を経る。

1. グラフ G0 に対して、C_i を選び、{v, w, C_i}と C_i の隣接点を縮約する。(G1)
2. G1 の independent domination number を求める。(k1)
3. さらに、ある C_j を縮約したグラフの independent domination number と比較する。(k2)

4. k1 > k2 ならば、G1 から C_j を縮約する。(G2)

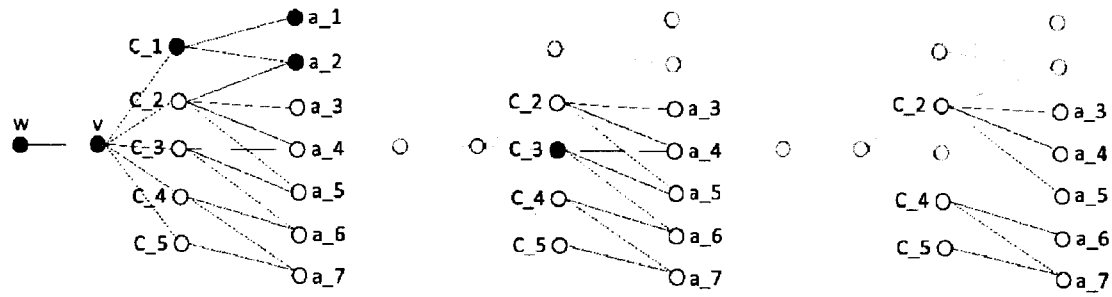


図 7　domination number を与える strategy (G0, G1, G2)

図 7 について、SAS programming は次のような例になる。

(1) delete from edge where var1 = w or var2 = v or var1 = C_1 or var2 = a_1 or var2 = a_2;

(2) delete from edge where var1 = C_3;

(3) if k1 > k2 then output (2); else output (1);

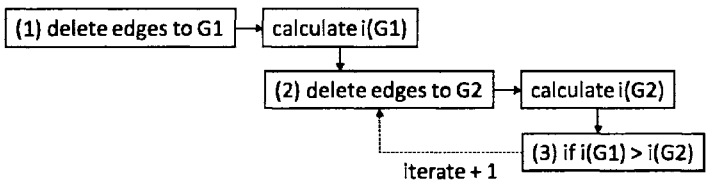


図 8　SAS programming の structure

これらにより、最初の C_i について C_j の集合が決まり、これを差集合として、任意の i における最小の被覆集合が導かれる。

リスクマネジメントの原則の留保

ISO によるリスクマネジメントの原則 [8] に鑑み、Domination model を標準とする。model の適用は、リスクへの対応に処する資質を制限し、全ての要素への分配や、未知なる創造性への投資を促進する。組織体を俯瞰することを可能にし、組織過程の支柱である。全体の同時的な把握により、意思決定に付加価値を与える。仮説や不確実性に対して、全てをカバーする。構造的に安定した戦略化を可能にする。常時存在からの発信を通じ、その中から最良の情報に基づく。model は任意に適合可能である。要素間の関係を記し、コミュニケーション過程を説明する。model 化により透明性と、包括性を保つ。全ての要素の資質に対する近接からその変化を動的、段階的、反映的に扱う。常時性により継続的な改善と増進を可能にする。互いへの経過により、継続的かつ周期的に再評価する。一方で、組織におけるリスクマネジメントを一様化するものではない。

参考文献

[1] ISO/IEC Guide 73:2009 (2009). Risk management ? Vocabulary. International Organization for Standardization.

[2] ISO/DIS 31000 (2009). Risk management ? Principles and guidelines on implementation. International Organization for Standardization.

[3] Reinhard Diestel (2010) Graph Theory Fourth Edition. Springer.

[4] Wu J. and Li H., "On calculating connected dominating set for efficient routing in ad hoc wireless networks", Proceedings of the 3rd International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications, (1999) ACM, pp. 7-14.

[5] M. R. Garey and D. S. Johnson (1979) Computers and Intractability - A Guide to the Theory of NP-Completeness. Freeman, San Francisco.

[6] Alan A. Bertossi, "Dominating sets for split and bipartite graphs", Information Processing Letters, 19 (1984), pp. 37-40.

[7] Harary F (1969) Graph theory. Addison-Wesley, Reading.

[8] "Committee Draft of ISO 31000 Risk management". International Organization for Standardization. 2007-06-15.

医薬品開発

生物学的同等性試験における例数設計：正確，近似と漸近

張 方紅・安藤 英一
グラクソ・スミスクライン（株）バイオメディカルデータサイエンス部

Sample size for bioequivalence trials: Exact, approximate and asymptotic methods

Fanghong Zhang and Hidekazu Ando
Biomedical Data Science Department,
GlaxoSmithKline K.K.

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨：

生物学的同等性試験の症例数設計について，Two One-Sided Test法における検出力の計算に焦点をあてる．正規分布，t分布，非心t分布，またはOwenのQ関数を用いた検出力の算出方法を説明する．POWERプロシジャの指定法を例示する．

キーワード：生物学同等性試験，TOST，POWERプロシジャ，OwenのQ関数，t分布

背景(1/2)

- 日，米，欧で生物学的同等性(BE)試験のガイドラインが存在
- 同等性の統計的な判定方法は古くから研究されてきた
  - 信頼区間法：Westlake (1972)
  - 検定法：Shuirmann (1987)のTOST (Two One-Sided Test)
- 症例数設計の方法は多く存在し，SASで利用されている手法を含め，初心者にとっては混乱しやすい
- 複数のSAS指定方法が存在

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

背景(2/2)：手法分類

- 正確法：OwenのQ関数を利用して2変量非心tの分布の確率を計算
  - POWERプロシジャ
- 近似法：1変量で2変量の分布の確率を近似する
  - 中心t分布を利用
    - 近似検出力：浜田・臨床評価研究会(2005)；
    - 近似症例数：Chow and Wang (2001)
  - 非心t分布を利用：Julious (2004)の式(59)
- 漸近法：漸近正規分布を利用，近似症例数
  - Julious (2004)の式 (60), (62)

目的

- TOST法における検出力を計算する数理を整理し，SASでQ関数を利用する正確法と複数に存在する近似法，および漸近正規分布を利用する漸近法を説明する
- POWERプロシジャの指定法を紹介し，例示する

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

後発医薬品の生物学的同等性試験ガイドライン

- 試験計画：クロスオーバー
- 評価PKパラメータ：AUCとCmax
- 統計解析：対数変換して，90%信頼区間，または有意水準5%の2つの片側検定（two one-sided tests, TOST）で評価
- 許容域：log(0.8)～log(1.25), -0.223～0.223

クロスオーバー試験：記号

- X=log(AUC)，または，X=log(Cmax)
- μ：母平均，μ=EX
- 添え字TとR：試験製剤と標準製剤を表す
- Δ：許容限界値

群: i	被験者: j	時期: k		差
1: TR	1	$x_{T1}(x_{111})$	$x_{R1}(x_{112})$	X_{T1}-X_{R1}
2: RT	2	$x_{R2}(x_{221})$	$x_{T2}(x_{222})$	X_{R2}-X_{T2}

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

薬剤効果の検定：t-test

$$t = \frac{\bar{d}_{1\bullet} - \bar{d}_{2\bullet}}{\sqrt{\frac{2}{n}\hat{\sigma}_d^2}} = \frac{\bar{X}_{T\bullet} - \bar{X}_{R\bullet}}{\sqrt{\frac{2\hat{\sigma}_w^2}{N}}} \sim t(N-2)$$

$$\frac{X_{ij1} - X_{ij2}}{2} = d_{ij}^{*} \qquad \sigma_d^2 = Var(d_{ij}^{*})$$

$$\sigma_d^2 = \frac{\sigma_w^2}{2} \qquad \sigma_w^2 = \frac{1}{2}\sigma_D^2$$

$$\sigma_D^2 = Var(X_T - X_R)$$

N=2n, nは1群における症例数

Chow and wang (2001)

クロスオーバー試験：データ

図表8.4 最大血中濃度C_{max}「SASによる実験データの解析」

A: 群 i	R: 被験者 j	B: 実験時期 k			
		B_1: 第1回目		B_2: 第2回目	
A_1	R_1 R_2 R_3 R_4 R_5	C_1 新製品	211 318 459 399 316	C_2 従来品	418 319 580 347 303
A_2	R_6 R_7 R_8 R_9 R_{10}	C_2 従来品	304 428 588 370 317	C_1 新製品	465 397 316 325 302

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

σ_w^2 in SAS output

proc mixed; model cmax=a b c; random r; ods output covparms=cov ;	Covariance Parameter Estimates Cov Parm Estimate r 1080.23 Residual 8836.25

model cmax=a b c;
repeated c/sub=r type=CS R;
ods output covparms=cov ;
$\sigma_w^2 = \frac{1}{2}Var(y_1 - y_2)$
	Covariance Parameter Estimates
	CovParm Subject Estimate
	CS r 1080.23
	Residual 8836.25

model cmax=a b c;
repeated c/sub=r type=UN R;
ods output covparms=cov;
	Cov Parm Subject Estimate
	UN(1,1) r 6781.90
	UN(2,1) r 1080.22
	UN(2,2) r 13051
	(6781.90+13051-2×1080.22)/2

σ_W^2

同等性仮説

- 帰無仮説1: $H_{01}: \mu_T - \mu_R \leqq -\Delta$
- 対立仮説1: $H_{11}: \mu_T - \mu_R > -\Delta$

- 帰無仮説2: $H_{02}: \mu_T - \mu_R \geqq \Delta$
- 対立仮説2: $H_{12}: \mu_T - \mu_R < \Delta$

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

TOST: 検定統計量

$$\frac{H_{01}: -\Delta \qquad H_{02}: \Delta}{\mu_T - \mu_R}$$

$$T_L = \frac{\bar{X}_T - \bar{X}_R + \Delta}{\hat{\sigma}_W \sqrt{\frac{2}{N}}} \qquad T_U = \frac{\bar{X}_T - \bar{X}_R - \Delta}{\hat{\sigma}_W \sqrt{\frac{2}{N}}}$$

$\bar{X}_T, \bar{X}_R$: 試験製剤と標準製剤の標本平均

$X_{T1}, X_{R1}, \ldots, X_{TN}, X_{RN}$: N人の被験者に対する観測値

TOST：有意水準

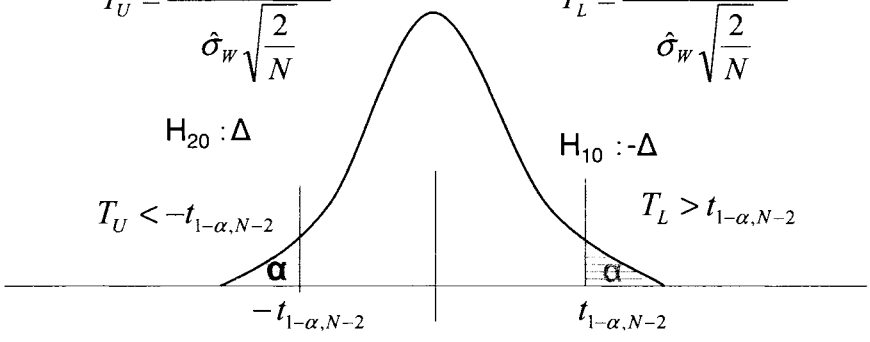
$$T_U = \frac{\bar{X}_T - \bar{X}_R - \Delta}{\hat{\sigma}_W \sqrt{\frac{2}{N}}} \qquad t(N-2) \qquad T_L = \frac{\bar{X}_T - \bar{X}_R + \Delta}{\hat{\sigma}_W \sqrt{\frac{2}{N}}}$$

$H_{20} : \Delta$

$H_{10} : -\Delta$

$T_U < -t_{1-\alpha,N-2}$

$T_L > t_{1-\alpha,N-2}$



TOST：検出力

対立仮説$-\Delta \leqq \mu_T - \mu_R \leqq \Delta$の下で

$$T_L = \frac{\bar{X}_T - \bar{X}_R + \Delta}{\hat{\sigma}_W \sqrt{\frac{2}{N}}} \qquad T_U = \frac{\bar{X}_T - \bar{X}_R - \Delta}{\hat{\sigma}_W \sqrt{\frac{2}{N}}}$$

$$Power = P\{T_L > t_{1-\alpha,N-2} \text{ and } T_U < -t_{1-\alpha,N-2} \mid \mu_T - \mu_R\}$$

検出力：1変量へ変形

- 対立仮説$-\Delta \leqq \mu_T - \mu_R \leqq \Delta$の下で

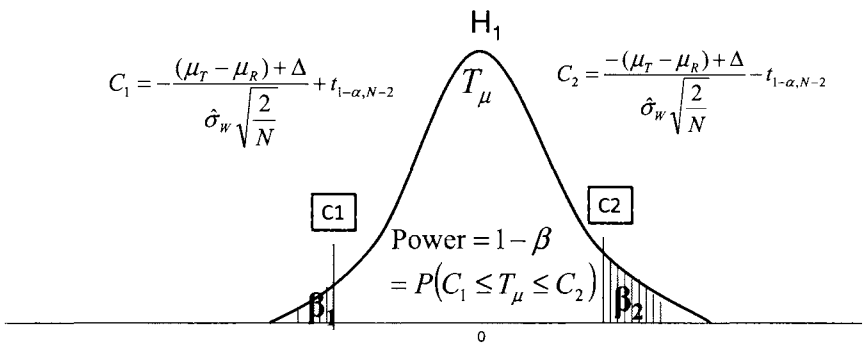
$$T_\mu = \frac{\overline{X}_T - \overline{X}_R - (\mu_T - \mu_R)}{\hat{\sigma}_W \sqrt{\frac{2}{N}}} \sim t(N-2)$$

$$Power = P\left\{ t_{1-\alpha,N-2} - \frac{\Delta + (\mu_T - \mu_R)}{\hat{\sigma}_W \sqrt{\frac{2}{N}}} < T_\mu < -t_{1-\alpha,N-2} + \frac{\Delta - (\mu_T - \mu_R)}{\hat{\sigma}_W \sqrt{\frac{2}{N}}} \right\}$$

矢船(2000), Chow and Wang (2001)

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

検出力



近似検出力：中心t分布

$$\text{Power} = P\{C_1 \le T_\mu \le C_2\}$$
$$\approx \text{PROBT}(C_2, N-2) - \text{PROBT}(C_1, N-2)$$

$$C_1 = -\frac{(\mu_T - \mu_R) + \Delta}{\hat{\sigma}_W\sqrt{\frac{2}{N}}} + t_{1-\alpha,N-2} \qquad C_2 = \frac{-(\mu_T - \mu_R) + \Delta}{\hat{\sigma}_W\sqrt{\frac{2}{N}}} - t_{1-\alpha,N-2}$$

浜田・臨床評価研究会(2005)

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

近似症例数：中心t分布

- $\mu_T = \mu_R$の場合, $\beta_1=\beta_2=\beta/2$, $-C_1=C_2 = t_{1-\beta/2, N-2}$

$$N = \frac{2\sigma_W^2 (t_{1-\alpha,N-2} + t_{1-\beta/2,N-2})^2}{\Delta^2}$$

- $\mu_T - \mu_R > 0$の場合, $\beta_1=0$, $\beta_2=\beta$, $C_2 = t_{1-\beta, N-2}$

$$N = \frac{2\sigma_W^2 (t_{1-\alpha,N-2} + t_{1-\beta,N-2})^2}{(\Delta - (\mu_T - \mu_R))^2}$$

Chow and Wang (2001)

症例数：漸近正規分布

- 大標本の場合，　$t_{1-\alpha,N-2} \approx z_{1-\alpha}$
- $\mu_T = \mu_R$

$$N = \frac{2\sigma_W^2 (z_{1-\alpha} + z_{1-\beta/2})^2}{\Delta^2}$$

- $\mu_T - \mu_R > 0$

$$N = \frac{2\sigma_W^2 (z_{1-\alpha} + z_{1-\beta})^2}{(\Delta - (\mu_T - \mu_R))^2}$$

Julious (2004)の式(62), (60)

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

近似検出力：中心t分布の問題点

- C_1とC_2が確率変数$\hat{\sigma}_W$を含んでいるため，定数ではない．中心t分布は，$\hat{\sigma}_W$の変動を考慮していない．

$$\text{Power} = P\{C_1 \le T_\mu \le C_2\}$$
$$\approx \text{PROBT}(C_2, N-2) - \text{PROBT}(C_1, N-2)$$

$$C_1 = -\frac{(\mu_T - \mu_R) + \Delta}{\hat{\sigma}_W \sqrt{\frac{2}{N}}} + t_{1-\alpha,N-2} \qquad C_2 = \frac{-(\mu_T - \mu_R) + \Delta}{\hat{\sigma}_W \sqrt{\frac{2}{N}}} - t_{1-\alpha,N-2}$$

検出力：非心t分布

対立仮説$-\Delta \leqq \mu_T - \mu_R \leqq \Delta$の下で

$$T_L = \frac{\bar{X}_T - \bar{X}_R + \Delta}{\hat{\sigma}_W \sqrt{\frac{2}{N}}} \sim t(N-2, \delta_1) \qquad T_U = \frac{\bar{X}_T - \bar{X}_R - \Delta}{\hat{\sigma}_W \sqrt{\frac{2}{N}}} \sim t(N-2, \delta_2)$$

$$\delta_1 = \frac{\mu_T - \mu_R + \Delta}{\sqrt{\frac{2}{N}}\sigma_W} \qquad \delta_2 = \frac{\mu_T - \mu_R - \Delta}{\sqrt{\frac{2}{N}}\sigma_W}$$

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

近似検出力

- Bonferroniの不等式

$$\begin{aligned}\text{Power} &= P\{A \cap B \mid \mu_T - \mu_R\} \\ &= 1 - P\{\bar{A} \cup \bar{B} \mid \mu_T - \mu_R\} \\ &\geq 1 - P(\bar{A} \mid \mu_T - \mu_R) - P(\bar{B} \mid \mu_T - \mu_R)\end{aligned}$$

$$A = \{T_L > t_{1-\alpha, N-2}\},\ B = \{T_U < -t_{1-\alpha, N-2}\}$$

$$T_L \sim t(N-2, \delta_1) \qquad T_U \sim t(N-2, \delta_2)$$

近似検出力

Power ≈ Julious (2004) の式(59)

$$= \mathrm{PROBT}(-t_{1-\alpha,N-2}, N-2, \delta_2) - \mathrm{PROBT}(t_{1-\alpha,N-2}, N-2, \delta_1)$$

Power ≈

$$1 - \mathrm{PROBT}(t_{1-\alpha,N-2}, N-2, \delta_1) - \mathrm{PROBT}(t_{1-\alpha,N-2}, N-2, -\delta_2)$$

この近似法が正確法より小さい検出力を算出
症例数を大きい方で近似

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

正確な検出力

$$T_L = \frac{Z+\delta_1}{\sqrt{\frac{S^2}{N-2}}} \quad T_U = \frac{Z+\delta_2}{\sqrt{\frac{S^2}{N-2}}} \quad Z \sim N(0,1), \quad S^2 \sim \chi^2(N-2)$$

Sun (2010)

$$\begin{aligned}\text{Power} &= P\{T_L > t_{1-\alpha,N-2}, T_U < -t_{1-\alpha,N-2} \mid \mu_T - \mu_R\} \\ &= \iint_D f(z,s)\,dz\,ds \qquad D = \{T_L > t_{1-\alpha,N-2}, T_U < -t_{1-\alpha,N-2}\} \\ &= Q_{N-2}(-t_{1-\alpha,N-2}, \delta_2; 0, R) - Q_{N-2}(t_{1-\alpha,N-2}, \delta_1; 0, R)\end{aligned}$$

- 逐次積分で重積分を計算

Owen' Q関数

$$\text{Power} = P\{T_L > t_{1-\alpha,\nu}, T_U < -t_{1-\alpha,\nu} \mid \mu_T - \mu_R\}$$
$$= Q_\nu(-t_{1-\alpha,\nu}, \delta_2; 0, R) - Q_\nu(t_{1-\alpha,\nu}, \delta_1; 0, R)$$

$$Q_\nu(t, \delta; 0, R) = P\left\{Z < \frac{t}{\sqrt{\nu}} S - \delta, \mathrm{S} \le \mathrm{R}\right\} = \int_0^R \Phi(\frac{t}{\sqrt{\nu}} s) f(s) ds$$

$$\mathrm{S} \sim f(s), \mathrm{S}^2 \sim \chi^2(\nu)$$

R：以下sに関する方程式の解

$$\frac{t_{1-\alpha,\nu}}{\sqrt{\nu}} s - \delta_1 = \frac{-t_{1-\alpha,\nu}}{\sqrt{\nu}} s - \delta_2$$

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

症例数設計：2つのデザイン

- 1標本デザイン：時期効果を無視，自由度N-1
- クロスオーバ：時期効果を考慮，自由度N-2

Julious (2004), 2.2.1節

POWERプロシジャ指定法：1標本(1/3)

proc power; $\sigma_D^2 = Var(X_T) - 2Cov(X_T, X_R) + Var(X_R)$
pairedmeans test=equiv_diff dist=normal
lower = log(0.8)
upper = log(1.25)
alpha = 0.05
pairedmeans=試験製剤平均｜標準製剤平均
pairedstddevs=試験製剤標準偏差｜標準製剤標準偏差
corr = 試験製剤と標準製剤間の相関係数
npairs = 総症例数
power = 検出力;
run;

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

POWERプロシジャ指定法：1標本(2/3)

proc power;
pairedmeans test=equiv_diff dist=normal
lower = log(0.8)
upper = log(1.25)
alpha = 0.05
meandiff=平均値の差
stddev=σ_W
corr =0
npairs = 総症例数
power = 検出力;
run;

$$Var(X_T) = Var(X_R) = \sigma_W^2$$

$$Cov(X_T, X_R) = 0$$

$$\sigma_D^2 = Var(X_T) + Var(X_R) = 2\sigma_W^2$$

POWERプロシジャ指定法：1標本(3/3)

```
proc power;
 pairedmeans test=equiv_ratio=lognormal
 lower    = 0.8
 upper    = 1.25
 alpha    = 0.05
 meanration=幾何平均値の比
 cv=被験者内変動係数
 corr     =0
 npairs   = 総症例数
 power    = 検出力;
run;
```

$$CV = \sqrt{e^{\sigma_W^2} - 1} \approx \sigma_W$$

浜田・安藤 (2006) p.37-38

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

POWERプロシジャ指定法：クロスオーバー

```
proc power;
 twosamplemeans test=equiv_diff dist=normal
 lower    = log(0.8)
 upper    = log(1.25)
 alpha    = 0.05
 meandiff=平均値の差
 stddev=σd
 ntotal= 総症例数
 power    = 検出力;
run;
```

$$\sigma_d^2 = Var\left(\frac{X_T - X_R}{2}\right) = \frac{\sigma_W^2}{2}$$

sun (2010)の Example 3

参考文献 (1/2)

- 小川幸男 (1997),生物学的同等性試験における例数設計. KR 研究会
- 高橋行雄, 大橋靖雄, 芳賀敏郎(1989). SASによる実験データの解析. 第8章. 東京大学出版社
- 浜田知久馬・安藤英一 (2006), POWERプロシジャによる症例数設計, SASユーザー会総会
- 浜田知久馬/監修・臨床評価研究会(ACE)基礎解析分科会/執筆 (2005), 実用SAS生物統計ハンドブック,サイエンティスト社
- 矢船明史 (2000),生物学的同等性試験における信頼区間に基づく例数設計について, 臨床薬理, 31(6)

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

参考文献 (2/2)

- Chow Shein-Chung and Wang Hansheng (2001), On Sample Size Calculation in Bioequivalence Trials. Journal of Pharmacokinetics and Pharmacodynamics, Vol. 28, No. 2
- Julious, Steven A. 2004. Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data. Statistics in Medicine, 23:1921-1986.
- Schuirman, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics 15, 657-680.
- Sun Peng (2010). Using SAS Proc Power to Perform Model-based Power Analysis for Clinical Pharmacology Studies. PharmaSUG2010 - Paper SP05
- Westlake, W. J. (1972). Use of confidence intervals in the analysis of comparative bioavailability trials. Journal of Pharmaceutical Sciences 61, 1340-1341.

イヌテレメトリー試験のデザインと統計解析法に関するシミュレーション検討

橋本 敏夫，中西 展大，河口 裕，武内 喜茂
田辺三菱製薬株式会社　研究本部　研究企画部

Monte Carlo simulation to assess the study design and statistical analysis method in the dog telemetry study

Toshio Hashimoto, Nobuhiro Nakanishi, Yutaka Kawaguchi and Yoshishige Takeuchi
Research Strategy & Planning Department, Mitsubishi Tanabe Pharma Corporation

要旨

医薬品の非臨床心血管系安全性評価の一環として，無麻酔無拘束下テレメトリー試験で，潜在的心毒性の指標であるQT/QTc間隔や血圧，心拍数などへの影響が評価される．テレメトリー試験では，4群×4期のクロスオーバーデザイン（Williamsラテン方格法，Peace用量漸増法）や用量漸増法などが用いられている．

SAS Mixedプロシジャを用いたモンテカルロシミュレーションにより，群間差の推定値と推定精度を算出し，それぞれの試験デザインにおける変動要因や統計解析法が群間比較に及ぼす影響を評価した．

試験デザインの検討では，Williamsラテン方格法が安定した推定精度を提供すること，およびPeace用量漸増法は推定精度が若干低下するが，群間差の推定への時期効果の介入を防止できることを確認した．

投与前値を共変量とした共分散分析の検討により，Williamsラテン方格法では，投与前値の調整が不要と考えられることと，用量漸増法では，投与前値を含む共分散分析によっても，時期効果による影響を十分に調整できない可能性を示した．また，欠測の補充方法と推定精度の関係について，複数の補充パターンを，シミュレーションで定量的に評価した．

多様な条件でのシミュレーションを行った本検討の成果は，QTcのみではなく，テレメトリー試験全般への適用が可能である．

キーワード：テレメトリー試験，クロスオーバーデザイン，Williamsラテン方格法，Peace用量漸増法，
用量漸増法，　共分散分析，欠測と補充

1．はじめに

非臨床試験におけるテレメトリー試験では，イヌやサルを対象にして，無麻酔無拘束下でQT/QTc間隔や血圧，心拍数などへの影響が評価される．日本製薬工業協会　統計DM部会のタスクフォースは，非臨床試験のデザインやQT補正の方法などを含む「QT延長の統計解析に関する解説書」（2007年）を作成したが，具体的なデータ解析法には言及されていない．

第5回日本安全性薬理研究会（2014年）において，テレメトリー試験の統計解析に関する特別セッション

が企画され，試験デザインとデータ解析法の標準化に関して議論された．第2期医薬安全性研究会安全性薬理チームは，日本安全性薬理研究会の活動を支援するとともに，統計的な課題についての継続検討を行い，第14回定例会（2014年）において，クロスオーバー試験に関する基礎的な解説や統計的な課題検討の結果を報告し，出席者と討論した．

我々は，これらの活動の一環として，少数例で実施されるテレメトリー試験の試験デザインと統計解析法に関する検討を行ったので，以下に報告する．シミュレーション検討の前半部分では，試験デザインと統計解析法の特性について，基礎的な事項の確認を行った．後半部分では，投与前値を共変量とした共分散分析の性能を評価し，欠測が発生した際の補充法が群間比較に与える影響を評価した．

2．シミュレーションの方法

2.1. 試験デザイン

シミュレーションで検討した試験デザインを図 1 に示した．Williams ラテン方格法および Peace 用量漸増法（以下，Peace 漸増法と略す）はいずれもクロスオーバーデザインであり，前者は各用量群の前後の時期の群の配置がバランスするように工夫され，後者は動物ごとに低用量から順に投与されるように工夫されている．用量漸増法は対照群から投与を開始し，低用量から順に投与する試験デザインである．

今回の検討は，動物 4 頭で 4 用量を評価する場合のシミュレーションとした．

図1．試験デザイン

動物番号	Williamsラテン方格法			
	1期	2期	3期	4期
1	1	2	3	4
2	2	4	1	3
3	3	1	4	2
4	4	3	2	1

動物番号	Peace用量漸増法			
	1期	2期	3期	4期
1	1	2	3	4
2	2	1	3	4
3	2	3	1	4
4	2	3	4	1

動物番号	用量漸増法			
	1期	2期	3期	4期
1	1	2	3	4
2	1	2	3	4
3	1	2	3	4
4	1	2	3	4

1:対照群 2:低用量 3:中用量 4:高用量

2.2. 統計モデル

以下のモデルにより発生させたデータを統計解析に使用した．

$$Y_{ijk} = \mu + dose_i + period_j + animal_k + e_{ijk}$$

μ : 総平均

$dose_i$: 投与量の効果 (i=1,2,3,4)

$period_j$: 時期効果 (j=1,2,3,4)

$animal_k$: 動物間変動 (k=1,2,3,4) $\sim$ N (0, σ_{animal})

e_{ijk} : 誤差変動 $\sim$ N (0, σ_e)

2.3. シミュレーションの条件

シミュレーションにおける固定効果と変動の大きさは，Williams ラテン方格法で実施されたテレメトリー試験のデータ解析結果（図 2）を参考に設定した．図 2 に Mixed プロシジャで解析して得られた最小 2 乗平均の差（図 2 上段），および共分散パラメータの推定値から求めた標準偏差（図 2 下段）を示した．図の左側は 30 分ごとの QTc の解析結果であり，右側は Sivarajah ら(2010)が提案したスーパーインターバルにしたがって，24 時間を 3 区間（1 : 2- 6 時間, 2 :7-14 時間, 3 : 14-24 時間）に分割した場合の解析結果である．

この結果を参考にして，各用量の母平均を{225, 230, 235, 250}，すなわち μ=225, $dose_i$={0, 5, 10, 25}とした．動物間変動および誤差変動は σ_{animal} = 5, σ_e = 5 に設定した．

時期効果については，日間変動として傾向的な偏りが介入した状況を想定して $period_j$={-4.5, -1.5, 1.5, 4.5}とした．

さらに，各要因の影響評価のために，QT延長作用がない帰無仮説条件 $dose_i$={0, 0, 0, 0}，時期効果がない条件 $period_j$={0, 0, 0, 0}，および動物間変動と誤差変動を変化させた条件（σ_{animal} =15, σ_e = 5），(σ_{animal} =5, σ_e = 1.67)についても検討した．

図2 シミュレーション条件の設定根拠

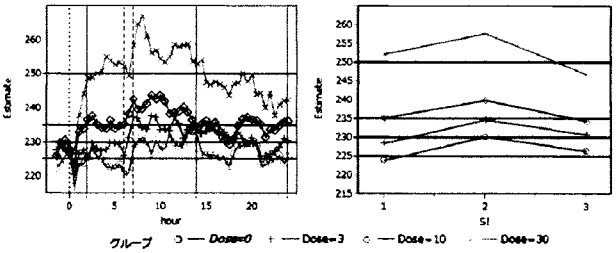


Fig2A Lsmeanの推移

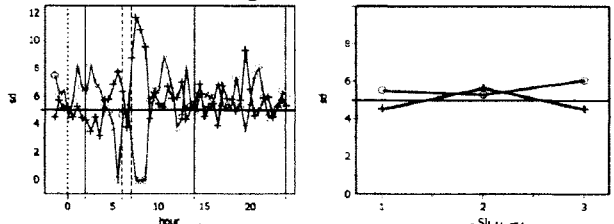


Fig2B σ_{animal}(○)， σ_e(+)の推移

2.4. シミュレーションの方法

2.2.統計モデルに記載した式に基づき，SAS9.2 データステップによりシミュレーションデータを作成した．各条件 10000 試験分のデータセットについて，以下の Mixed プロシジャで解析した．なお，多群比較においては Dunnett 検定などの多重比較法が用いられるが，本シミュレーションでは多重性の調整は考慮しなかった．

表 1-1．データ解析プログラム

Williams ラテン方格法，Peace 漸増法	用量漸増法
proc mixed data=dataset ; class animal period dose ; model y=dose period / ddfm=kenwardroger ; random animal ; lsmeans dose / pdiff=control('1') tdiff cl ; run ;	proc mixed data=dataset ; class animal dose ; model y=dose / ddfm=kenwardroger ; random animal ; lsmeans dose / pdiff=control('1') tdiff cl ; run ;

2.5. シミュレーション結果の評価

Williams ラテン方格法についての Mixed プロシジャの出力例を表 1-2 に示す．試験デザインや変動要因の影響等について，10000 回のシミュレーションの要約統計量により評価した．なお，評価には以下の指標を用いた．

1) 推定値（Estimate） ：各投与量と対照群の差の最小二乗平均（LSMEAN）
2) 推定精度（信頼区間幅） ：各投与量と対照群の差の推定値の両側 95%信頼区間の片幅（上限－推定値）
3) Pr > |t| ：対照群との有意差
4) 共分散パラメータ ：V_{animal}, $V_{residual}$

表 1-2. Mixed プロシジャの出力例（Williams ラテン方格法）

分類変数の水準の情報

分類	水準	値
animal	4	1 2 3 4
period	4	1 2 3 4
dose	4	1 2 3 4

共分散パラメータ推定値

共分散パラメータ	推定値
animal	24.3436
Residual	25.1037

固定効果の Type 3 検定

効果	分子の自由度	分母の自由度	F 値	Pr > F
dose	3	6	28.85	0.0006
period	3	6	1.85	0.2386

最小 2 乗平均の差

効果	dose	_dose	推定値	標準誤差	自由度	t 値	Pr > |t|	アルファ	下限	上限
dose	2	1	9.0983	3.5429	6	2.57	0.0424	0.05	0.4292	17.7674
dose	3	1	13.3405	3.5429	6	3.77	0.0093	0.05	4.6714	22.0096
dose	4	1	31.9842	3.5429	6	9.03	0.0001	0.05	23.3151	40.6533

3．試験デザインに関するシミュレーション

3.1. 推定値と推定精度（表 2）

Williams ラテン方格法，Peace 漸増法および用量漸増法における，各投与量の作用の推定値と推定精度について，10000 回のシミュレーションの平均値を表 2 に示した．

単調増加する時期効果を仮定したシミュレーションでは，用量漸増法は，推定値に時期効果がバイアスとして介入しているため，各用量の作用が過大に評価されていた．一方，Williams ラテン方格法および Peace 漸増法は，いずれの条件においても，各用量の作用を偏りなく推定していた．推定精度は，Peace 漸増法は Williams ラテン方格法よりも低かった（信頼区間幅が広かった）．表には示していないが，Peace 漸増法の推定値は，低用量と高用量でばらつきが大きくなる傾向を認めた．

時期効果がないと仮定したモデルでは，いずれのデザインにおいても各用量の作用の推定値は，偏りなく推定されていた．推定精度は，用量漸増法＞Williams ラテン方格法＞Peace 漸増法の順であった．

表2 試験デザインと推定値，推定精度

dose $_i$	試験デザイン	σ_{animal}	σ_e	period$_j$ { -4.5, -1.5, 1.5, 4.5 }				period$_j$ { 0, 0, 0, 0 }			
				推定値(Estimate)			推定精度 (信頼区間幅)	推定値(Estimate)			推定精度 (信頼区間幅)
				低用量	中用量	高用量		低用量	中用量	高用量	
{ 0, 5, 10, 25 }	Williams ラテン方格	5	5	5.01	10.02	25.03	8.13	5.03	10.01	24.99	8.13
		15	5	5.00	10.02	24.98	8.32	5.05	10.06	25.01	8.34
		5	1.67	4.99	9.99	24.99	2.77	4.98	10.00	25.00	2.77
	Peace 漸増法	5	5	5.02	9.97	24.96	9.87	4.98	9.99	24.99	9.79
		15	5	4.96	9.97	24.96	10.04	5.02	9.98	25.06	10.07
		5	1.67	4.99	9.99	24.99	3.34	5.00	10.01	25.01	3.36
	用量 漸増法	5	5	7.98	15.96	33.94	7.70	5.09	10.00	25.04	7.68
		15	5	7.95	16.00	34.01	7.79	5.00	9.97	24.95	7.80
		5	1.67	7.99	16.00	33.98	2.59	5.00	10.00	25.00	2.60
{ 0, 0, 0, 0 }	Williams ラテン方格	5	5	0.01	0.02	0.03	8.13	0.03	0.01	-0.01	8.13
		15	5	0.00	0.02	-0.02	8.32	0.05	0.06	0.01	8.34
		5	1.67	-0.01	-0.01	-0.01	2.77	-0.02	0.00	0.00	2.77
	Peace 漸増法	5	5	0.02	-0.03	-0.04	9.87	-0.02	-0.01	-0.01	9.79
		15	5	-0.04	-0.03	-0.04	10.04	0.02	-0.02	0.06	10.07
		5	1.67	-0.01	-0.01	-0.01	3.34	0.00	0.01	0.01	3.36
	用量 漸増法	5	5	2.98	5.96	8.94	7.70	0.09	0.00	0.04	7.68
		15	5	2.95	6.00	9.01	7.79	0.00	-0.03	-0.05	7.80
		5	1.67	2.99	6.00	8.98	2.59	0.00	0.00	0.00	2.60

10000回の平均値

3.2. 共分散パラメータ（表 3）

共分散パラメータの推定値についての集計結果を表 3 に示した．V_{animal} と $V_{residual}$ のメディアンは，それぞれ σ_{animal} と σ_e の設定値の二乗より若干小さな値となったが，平均値は σ^2_{animal} と σ^2_e に近似しており，不偏推定量を与えることを確認した．

表3 試験デザインと分散の推定値

dose $_i$	試験デザイン	σ_{animal}	σ_e	period$_j$ { -4.5, -1.5, 1.5, 4.5 }				period$_j$ { 0, 0, 0, 0 }			
				V_{animal}		$V_{residual}$		V_{animal}		$V_{residual}$	
				mean	median	mean	median	mean	median	mean	median
{ 0, 5, 10, 25 }	Williams ラテン方格	5	5	25.1	18.3	24.4	21.9	25.1	18.4	24.3	21.8
		15	5	222.9	174.7	25.1	22.5	223.5	177.1	25.2	22.6
		5	1.67	25.1	19.7	2.8	2.5	25.3	19.7	2.8	2.5
	Peace 漸増法	5	5	25.1	18.4	24.4	22.0	25.2	18.5	24.1	21.5
		15	5	223.0	174.7	25.0	22.5	223.5	177.3	25.1	22.4
		5	1.67	25.1	19.7	2.8	2.5	25.3	19.7	2.8	2.5
	用量 漸増法	5	5	25.1	18.3	24.7	22.9	25.1	18.5	24.5	22.8
		15	5	223.0	174.8	25.1	23.5	223.5	177.2	25.2	23.3
		5	1.67	25.1	19.7	2.8	2.6	25.3	19.7	2.8	2.6
{ 0, 0, 0, 0 }	Williams ラテン方格	5	5	25.1	18.3	24.4	21.9	25.1	18.4	24.3	21.8
		15	5	222.9	174.7	25.1	22.5	223.5	177.1	25.2	22.6
		5	1.67	25.1	19.7	2.8	2.5	25.3	19.7	2.8	2.5
	Peace 漸増法	5	5	25.1	18.4	24.4	22.0	25.2	18.5	24.1	21.5
		15	5	223.0	174.7	25.0	22.5	223.5	177.3	25.1	22.4
		5	1.67	25.1	19.7	2.8	2.5	25.3	19.7	2.8	2.5
	用量 漸増法	5	5	25.1	18.3	24.7	22.9	25.1	18.5	24.5	22.8
		15	5	223.0	174.8	25.1	23.5	223.5	177.2	25.2	23.3
		5	1.67	25.1	19.7	2.8	2.6	25.3	19.7	2.8	2.6

3.3. 試験デザインと検出力（表 4）

対照群と各用量で有意となった割合を表 4 に示した．左の列は，単調増加する時期効果を仮定した場合，右の列は時期効果を仮定しなかった場合の結果である．表の上段（dose $_i$ { 0, 5, 10, 25 }）は，各投与量（Δ =5, 10, 25）の検出力に相当する．いずれのデザインも動物間変動の大きさには影響を受けず，誤差変動

の大きさに影響を受けることがわかる．また，Williams ラテン方格法と，Peace 漸増法において，時期効果の有無は検出力に影響を与えなかった．

表の下段（dose$_i$ { 0, 0, 0, 0 }）は，帰無仮説条件下における第 1 種過誤率（危険率）に相当する．Williams ラテン方格法と，Peace 漸増法の危険率が 5%に維持されていることが確認された．一方，用量漸増法では，時期効果がある場合に，11.1%～100%と，条件により 5%を大きく上回った．

表4 対照群との有意差を認めた頻度

dose$_i$	試験デザイン	σ_{animal}	σ_e	period$_j$ { -4.5, -1.5, 1.5, 4.5 }			period$_j$ { 0, 0, 0, 0 }		
				低用量	中用量	高用量	低用量	中用量	高用量
{ 0, 5, 10, 25 }	Williams ラテン方格	5	5	23.1%	67.2%	100.0%	23.0%	67.5%	100.0%
		15	5	22.0%	65.5%	100.0%	21.8%	66.1%	100.0%
		5	1.67	93.7%	100.0%	100.0%	93.5%	100.0%	100.0%
	Peace 漸増法	5	5	17.0%	59.4%	99.8%	18.1%	59.7%	99.7%
		15	5	16.7%	57.7%	99.7%	16.3%	58.4%	99.7%
		5	1.67	83.3%	100.0%	100.0%	82.6%	100.0%	100.0%
	用量漸増法	5	5	55.1%	97.8%	100.0%	26.3%	72.0%	100.0%
		15	5	54.0%	97.6%	100.0%	24.5%	70.9%	100.0%
		5	1.67	100.0%	100.0%	100.0%	96.0%	100.0%	100.0%
{ 0, 0, 0, 0 }	Williams ラテン方格	5	5	5.1%	5.4%	5.1%	5.5%	5.1%	5.3%
		15	5	4.7%	4.8%	5.2%	5.3%	4.9%	5.2%
		5	1.67	5.1%	5.0%	5.4%	5.2%	4.7%	5.1%
	Peace 漸増法	5	5	5.1%	4.8%	5.1%	5.8%	5.3%	5.4%
		15	5	5.1%	5.2%	5.2%	4.8%	5.0%	5.2%
		5	1.67	4.9%	5.2%	5.1%	4.9%	4.8%	5.3%
	用量漸増法	5	5	12.4%	33.4%	62.4%	5.0%	5.3%	5.1%
		15	5	11.1%	33.1%	61.9%	5.5%	4.7%	5.0%
		5	1.67	62.0%	99.5%	100.0%	5.2%	4.7%	5.1%

上段:検出力, 下段:第1種の過誤率

3.4. 統計手法を誤用した場合の影響（表 5）

統計解析環境が整備されていない等の理由から，ラテン方格デザインで実施した試験結果を，一元配置分散分析で解析するという誤用が，まれに見受けられるようである．そこで，Williams ラテン方格法で実施された試験結果に一元配置型の解析を適用した場合を，シミュレーションで検討した．

表 5 に推定値，推定精度およびラテン方格分散分析で解析した場合の推定精度(表 2)との比（信頼区間幅の比）を示した．適切な解析法が採用されない場合には，推定値には影響しないが，推定精度が低下する．すなわち，一元配置型の解析を用いることにより，信頼区間幅は，時期効果ありで 1.5~3.5 倍，時期効果なしで 1.3~2.8 倍となった．その傾向は時期効果ありで強く，特に動物間変動と誤差変動が $\sigma_{animal} > \sigma_e$ の条件では，約 3 倍と推定精度が大きく低下することがわかる．

表5 Williamsラテン方格法で得られた試験結果に一元配置型の解析を適用した場合

dose$_i$	試験デザイン	σ_{animal}	σ_e	period$_j$ { -4.5, -1.5, 1.5, 4.5 }					period$_j$ { 0, 0, 0, 0 }				
				推定値(Estimate)			推定精度	表2との比	推定値(Estimate)			推定精度	表2との比
				低用量	中用量	高用量			低用量	中用量	高用量		
{ 0, 5, 10, 25 }	Williams ラテン方格	5	5	5.01	10.02	25.03	12.08	**1.5**	5.03	10.01	24.99	10.50	**1.3**
		15	5	5.00	10.02	24.98	23.58	**2.8**	5.05	10.06	25.01	22.75	**2.7**
		5	1.67	4.99	9.99	24.99	9.81	**3.5**	4.98	10.00	25.00	7.64	**2.8**
{ 0, 0, 0, 0 }	Williams ラテン方格	5	5	0.01	0.02	0.03	12.08	**1.5**	0.03	0.01	-0.01	10.50	**1.3**
		15	5	0.00	0.02	-0.02	23.58	**2.8**	0.05	0.06	0.01	22.75	**2.7**
		5	1.67	-0.01	-0.01	-0.01	9.81	**3.5**	-0.02	0.00	0.00	7.64	**2.8**

10000回の平均値

4．投与前値を共変量とする共分散分析の評価

投与前値を共変量とする共分散分析の必要性と性能を評価するために，Williams ラテン方格法と用量漸増法について，それぞれ，投与前値を考慮しない解析(Analysis 1)，投与前値を共変量とした共分散分析(Analysis 2)，および animal を変量効果に含めない共分散分析(Analysis 3)の 3 種の解析方法についてシミュレーションを実施した．投与前値は投与量の効果($dose_i$)を含まないデータとして，以下のモデルにより生成した．

投与前値（pre）：$X_{ijk} = \mu + period_j + animal_k + e'_{ijk}$

測定値　　：$Y_{ijk} = \mu + dose_i + period_j + animal_k + e_{ijk}$

シミュレーションに用いた SAS プログラムを表 6 に，シミュレーション結果を表 7 に示した．

表 6. 解析プログラム

Williams ラテン方格法	用量漸増法
Analysis 1　（前値を考慮しない解析）	Analysis 1　（前値を考慮しない解析）
proc mixed data=dataset ; class animal period dose ; model y = **dose period** / ddfm=kenwardroger ; random **animal** ; lsmeans dose / pdiff=control('1')　tdiff cl ; run ;	proc mixed data=dataset ; class animal dose ; model y=**dose** / ddfm=kenwardroger ; random **animal** ; lsmeans dose / pdiff=control('1') tdiff cl ; run ;
Analysis 2　（前値を含めた共分散分析）	Analysis 2　（前値を含めた共分散分析）
proc mixed data=dataset ; class animal period dose ; model y=**dose period pre** / ddfm=kenwardroger ; random **animal**; lsmeans dose / pdiff=control('1') tdiff cl ; run ;	proc mixed data=dataset; class animal dose; model y=**dose pre** / ddfm=kenwardroger ; random **animal** ; lsmeans dose / pdiff=control('1') tdiff cl ; run ;
Analysis 3　（period を除いた共分散分析）	Analysis 3　（animal を除いた共分散分析）
proc mixed data=dataset ; class animal period dose ; model y=**dose pre** / ddfm=kenwardroger ; random **animal** ; lsmeans dose / pdiff=control('1') tdiff cl ; run ;	proc mixed data=dataset ; class animal dose ; model y=**dose pre**/ ddfm=kenwardroger ; lsmeans dose / pdiff=control('1') tdiff cl ; run ;

太字・下線部はモデルに含めた要因である（“pre”は投与前値）．

表7 投与前値を共変量とした共分散分析

dose $_i$	試験デザイン	σ_{animal}	σ_e	Williamsラテン方格法				用量漸増法			
				推定値(Estimate)			推定精度	推定値(Estimate)			推定精度
				低用量	中用量	高用量		低用量	中用量	高用量	
{ 0, 5, 10, 25 }	Analysis 1	5	5	5.02	10.06	25.02	8.14	7.99	16.05	33.98	7.69
		15	5	4.96	9.96	24.99	8.26	8.00	15.94	33.94	7.76
		5	1.67	5.01	10.02	25.01	2.76	8.00	16.02	33.99	2.59
	Analysis 2	5	5	5.05	10.09	25.01	8.98	7.21	14.50	31.64	8.45
		15	5	4.97	9.93	25.01	10.18	6.02	11.99	27.98	9.43
		5	1.67	5.02	10.04	25.01	3.40	6.04	12.09	28.09	3.40
	Analysis 3	5	5	5.06	10.08	25.03	9.32	6.68	13.41	30.00	9.47
		15	5	4.97	9.92	25.00	9.77	5.49	10.92	26.39	10.41
		5	1.67	5.02	10.03	25.01	3.40	5.49	10.99	26.45	3.61
{ 0, 0, 0, 0 }	Analysis 1	5	5	0.02	0.06	0.02	8.14	2.99	6.05	8.98	7.69
		15	5	−0.04	−0.04	−0.01	8.26	3.00	5.94	8.94	7.76
		5	1.67	0.01	0.02	0.01	2.76	3.00	6.02	8.99	2.59
	Analysis 2	5	5	0.05	0.09	0.01	8.98	2.21	4.50	6.64	8.45
		15	5	−0.03	−0.07	0.01	10.18	1.02	1.99	2.98	9.43
		5	1.67	0.02	0.04	0.01	3.40	1.04	2.09	3.09	3.40
	Analysis 3	5	5	0.06	0.08	0.03	9.32	1.68	3.41	5.00	9.47
		15	5	−0.03	−0.08	0.00	9.77	0.49	0.92	1.39	10.41
		5	1.67	0.02	0.03	0.01	3.40	0.49	0.99	1.45	3.61

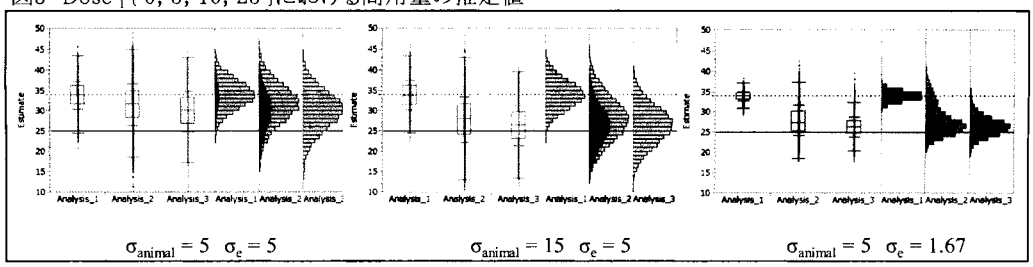
period $_j$ { −4.5, −1.5, 1.5, 4.5 }　　10000回の平均値

Williams ラテン方格法で実施された試験結果は，いずれの解析方法においても推定値に偏りが生じなかった．推定精度は，投与前値を考慮しない Analysis 1 で最も高かった．

用量漸増法で実施された試験結果については，投与前値を考慮しない Analysis 1 で，推定値は設定値より時期効果の分だけ大きくなるというバイアスが介入した．投与量と投与前値の 2 要因とした Analysis 3 では，σ_{animal} と σ_e との比が大きな条件で，推定値が設定値に近づいたが，Williams ラテン方格法のように時期効果の影響を完全に調整することはできなかった．Analysis 2 は，Analysis 1 と Analysis 3 の中間的な値を示し，時期効果の影響を調整しきれていなかった．

用量漸増法の dose $_i$ { 0, 5, 10, 25 }における高用量の推定値の分布を箱ひげ図とヒストグラムで示す（図3）．ヒストグラムの網掛けの部分は，Mixed プロシジャで animal の共分散の推定値が 0 と出力されたケースであり，Analysis1 に比べ投与前値を共変量とした共分散分析(Analysis 2)において頻度が高かった．Mixed プロシジャでは，指定した変量効果が誤差変動に比べて十分に小さいとみなせる場合には，その変量効果をモデルから外していた．このため，Analysis 2 の網掛け部分は，Analysis 3 と同様の解析結果となり，投与前値による共分散分析の推定値がばらついていた．

図3 Dose $_i$ { 0, 5, 10, 25 }における高用量の推定値



$\sigma_{animal} = 5$　$\sigma_e = 5$　　$\sigma_{animal} = 15$　$\sigma_e = 5$　　$\sigma_{animal} = 5$　$\sigma_e = 1.67$

5．欠測が生じた場合の補充の方法と推定精度

Williams ラテン方格法で，欠測が発生した場合の補完方法について，シミュレーションを行った．傾向的な時期効果が発生している条件(period$_j$ { -4.5, -1.5, 1.5, 4.5, 7.5 })を設定し，動物 3 の第 3 期の高用量が欠測する 7 つのパターンを検討した．検討したパターンを図 4 に，推定値と推定精度（比較のために欠測が発生しない状況でのシミュレーション結果を含めた）を表 8 に示した．

パターン①は欠測を補充しないケースである．パターン②～⑤は，欠測が発生した動物での再試験が可能なケースで，第 5 期に実験を追加している．パターン⑥，⑦は，欠測が発生した動物 3 が，第 3 期以降の実験に使用できなくなるため，第 4 期から動物 5 を追加するというケースである．

図4 欠測と補充のパターン

① n=15	1期	2期	3期	4期
動物1	1	2	3	4
動物2	2	4	1	3
動物3	3	1	×	2
動物4	4	3	2	1

② n=16	1期	2期	3期	4期	5期
動物1	1	2	3	4	
動物2	2	4	1	3	
動物3	3	1	×	2	4
動物4	4	3	2	1	

③ n=16	1期	2期	3期	4期	3'期
動物1	1	2	3	4	
動物2	2	4	1	3	
動物3	3	1	×	2 ←	4
動物4	4	3	2	1	

④ n=19	1期	2期	3期	4期	5期
動物1	1	2	3	4	3
動物2	2	4	1	3	1
動物3	3	1	×	2	4
動物4	4	3	2	1	2

⑤ n=17	1期	2期	3期	4期	5期
動物1	1	2	3	4	
動物2	2	4	1	3	1
動物3	3	1	×	2	4
動物4	4	3	2	1	

⑥ n=19	1期	2期	3期	4期	5期
動物1	1	2	3	4	3
動物2	2	4	1	3	1
動物3	3	1	×		
動物4	4	3	2	1	2
動物5				2	4

⑦ n=17	1期	2期	3期	4期	5期
動物1	1	2	3	4	
動物2	2	4	1	3	1
動物3	3	1	×		
動物4	4	3	2	1	
動物5				2	4

⊠ 動物3 第3期の高用量で欠測発生

1:対照群 2:低用量 3:中用量 4:高用量

表8 高用量の推定値と推定精度

欠測と補充パターン	dose$_i$ { 0, 5, 10, 25 }						dose$_i$ { 0, 0, 0, 0 }					
	推定値(Estimate)			推定精度			推定値(Estimate)			推定精度		
σ_{animal}	5	15	5	5	15	5	5	15	5	5	15	5
σ_e	5	5	1.67	5	5	1.67	5	5	1.67	5	5	1.67
欠測なし	**25.03**	**24.98**	**24.99**	**8.13**	**8.32**	**2.77**	**0.03**	**−0.02**	**−0.01**	**8.13**	**8.32**	**2.77**
①	25.02	25.00	25.00	9.62	9.96	3.32	0.02	0.00	0.00	9.62	9.96	3.32
②	25.02	25.00	25.00	9.62	9.96	3.32	0.02	0.00	0.00	9.62	9.96	3.32
③	26.50	26.50	26.51	8.45	8.64	3.75	1.50	1.50	1.51	8.45	8.64	3.75
④	24.99	25.03	25.00	7.59	7.67	2.55	−0.01	0.03	0.00	7.59	7.66	2.55
⑤	24.99	25.03	25.00	7.94	8.08	2.70	−0.01	0.03	0.00	7.94	8.08	2.70
⑥	24.97	25.01	25.00	8.11	8.61	2.87	−0.03	0.01	0.00	8.11	8.61	2.87
⑦	24.98	25.01	25.00	8.54	9.30	3.09	−0.02	0.01	0.00	8.54	9.30	3.09

period$_j$ { −4.5, −1.5, 1.5, 4.5, 7.5 }　　10000回の平均値

第 5 期に得られた高用量のデータを第 3 期とみなす，パターン③では，欠測が発生しない場合に比し推定値が大きく，時期効果によるバイアスが介入した．その他のパターンでは，欠測が発生しない場合と近似し，推定値へのバイアスを認めなかった．

欠測を補充しないパターン①では推定精度が低下し，信頼区間幅が 20%程度広くなった．第 5 期に 1 例だけを補充するパターン②は．補充なしのパターン①と同じ推定精度であった．第 5 期に 4 例を評価する

パターン④の推定精度は欠測なしよりも高く（信頼区間幅が狭く）なり，第5期に高用量と対照群の2例のみを評価するパターン⑤の推定精度は，欠測なしに近似した．

別動物で補充するケースでは，第5期に4例を評価するパターン⑥の推定精度が欠測なしの推定精度に近似し，第5期に2例のみを評価するパターン⑦の推定精度は，欠測なしに比べて低下し，信頼区間幅が10%程度広くなった．

6．まとめ

モンテカルロシミュレーションにより，動物4頭を使ったテレメトリー試験の試験デザインと統計解析法の特性を評価した．

Williams ラテン方格法は，時期効果の影響を適切に排除しており，群間差の推定も安定する優れた試験デザインであることを確認した．この試験デザインにおいては，共分散分析による投与前値の調整は不要と考えられた．

Peace 漸増法は，群間差の推定精度の面で Williams ラテン方格法よりも若干劣るが，群間差の推定への時期効果の介入を防ぐことができることを確認した．安全性を確保するなどで，各動物に低用量から順の投与が有益なケースでは，Peace 漸増法の選択が推奨される．

用量漸増法は，時期効果が群間差の推定に直接介入すること，その影響は，投与前値を共変量とした共分散分析によっても調整が不十分となる場合があることを示した．よって，非可逆的な毒性の発現が懸念されるなどで上記試験デザインが採用できず，やむを得ず本デザインで試験を行う場合は，時期効果を排除するための配慮が必要である．

SAS Mixed プロシジャは，少数例のクロスオーバー試験の解析においても，投与量の効果，動物間変動および誤差変動等で，偏りのない推定値を与えることを確認した．検出力および第1種の過誤の評価では，Williams ラテン方格法と Peace 漸増法は時期効果の影響を受けず第1種の過誤を5%に保持すること，および用量漸増法では時期効果がある場合，第1種の過誤を5%に維持できないことを確認した．クロスオーバー法で実施した試験結果を一元配置型で解析するという統計手法の誤用を行った場合には，推定精度が低下することを確認した．以上の検討により，テレメトリー試験の評価には，時期効果を考慮したクロスオーバーデザインの選択と，投与量の効果，時期効果，動物の変動を考慮した統計解析法の使用が重要であることを示した．

Williams ラテン方格法で欠測が生じた場合の補充に関して7つのパターンについてのシミュレーションを行った．欠測した群の1例を第5期に追加するパターン②，パターン③の選択を避けることを推奨する．試験計画の検討にあたっては，欠測の発生要因，実施可能性に加えて，推定値や推定精度への影響を考慮した検討が有用であり，多様な欠測・補充パターンに関する更なる検討も必要と考える．

7．おわりに

本検討にあたり，第2期医薬安全性研究会 安全性薬理チームの高橋 行雄氏，半田 淳氏，山田 雅之氏，福島 慎二氏，平田 篤由氏，板東 正博氏，金納 明宏氏に，多くのご助言をいただいた．非臨床試験の試験デザインの検討や適切な統計解析の使用には，非臨床研究者と生物統計の専門家の連携が必要と考える．更なる連携の強化と検討の深化に期待する．

8．文献

[1] 医薬品評価委員会 統計・DM 部会(2007), QT 延長の統計解析に関する解説書,日本製薬工業協会 医薬出版センター

[2] 板東 正博(2014)，覚醒テレメトリー犬における心血管系安全性評価 －試験デザイン，データ解析，統計手法の提案－，第 5 回日本安全性薬理研究会学術年会

[3] 古賀 正ら(2014)，安全性薬理コアバッテリーの心血管系に関する統計解析の現状，第 5 回日本安全性薬理研究会学術年会

[4] A. Sivarajah et. al. (2010), Cardiovascular safety assessments in the conscious telemetered dog of super-intervals to enhance statistical power, J Pharmacol. Toxicol method, 62 12–19

[5] M. Aylott et al.(2011), Review of the statistical analysis of the dog telemetry study, Pharmaceut. Statist., 10 236–249

以上

Model based LibraryによるLogical Checkの自動生成
- チェック仕様書作成業務の効率化 -

三木 悠吾
DOTインターナショナル株式会社 データサイエンス部

A Generation of Logical Check Codes by Model based Library For Productive Creation of Check Specifications

Yugo Miki
Data Science Dept, DOT international Co., Ltd.

要旨

医薬品開発において統計解析を実施するためには、データクリーニングが必須であるがこの業務は決して工数の少ないものではない。弊社では業務効率化のためにmodel based開発を応用したロジカルチェックの自動生成を試みた。

キーワード: model based、Data management、logical check、call execute

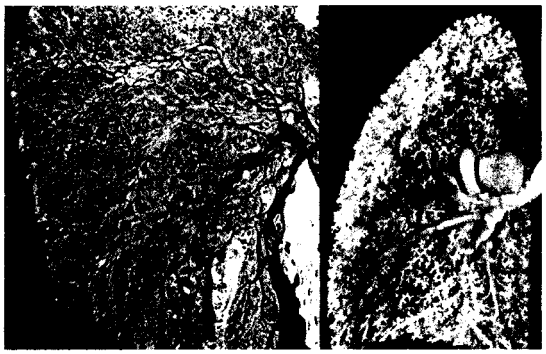
目次

Model based LibraryによるLogical Checkの自動生成と
チェック仕様書作成業務の効率化

PROPAGANDA

哲学 – *the Constructal Law* –

- コンストラクタル法則
  - *有限の流動系が時間の流れの中で存続するためには、その流動系の配置は、流動抵抗を低減するように進化しなければならない* *1



Flow systems in action: the delta of the Lena River in northern Siberia (left) and a cast of a human lung (right).

(A. Bejan)

*1 木村繁男 訳

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

哲学 – *the Constructal Law for Data* –

- 情報伝達における進化の方向性とは
  - 進化は効率的な方向へ進む
    - ex. 絵→象形文字→文字 （より明確に）
      会話→電話→衛星通信 （より遠くに）
      手紙→電子メール （より素早く）
      紙→本→電子媒体 （より多くの情報を）

社内における役割別SAS習得度

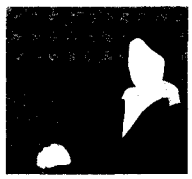
Monitor
SAS習得レベル：微

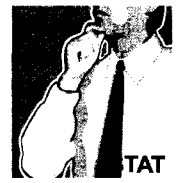
Data Manager
SAS習得レベル：中

Statician
SAS習得レベル：高

個人差はあるが基本的にSTATチームが高い
社内で使用するSAS組み込みシステムなどは社内SEでなくSAS programmer
の上級者が開発している現状

Check仕様書作成において





・エラーメッセージの作成
・チェックの妥当性確認

・チェック項目の策定
・CRAへ項目の確認
・STATへ委託
・疑義に回答する
・訂正が発生したら・・・

・Check Programの作成
・CSV or Test
・Check仕様書に疑義を上げる

DMの仕事量がすごく多い・・・・
疑義事項を挙げるプロセスだけでもけずれないか・・・？

仕様書に関する問題点

- 書き手の意思を読み手に確実に伝えることができない
  - 何かしらの齟齬が発生する可能性を否めない
- 「正しく」仕様書を書くコストは高くつく
  - SASに関しての知識
  - データベースに関する知識
  - 臨床試験に関する知識など・・・・
- 完全な仕様を求められる
  - 規制法の中なので仕方ない・・・
  - たとえcheckが500個であっても

➢ 必要なことではあるが、checkの品質を担保するには工数がかかりすぎる・・・・

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

モデルベース開発とは

- 組み込みシステムなどで発達してきた開発手法
- 直感的に理解しやすいモデルを仕様書とすることができる
- モデルベース開発の特徴
  - モデルによる仕様の表現・定義「実行可能な仕様書」
  - モデルのシミュレーションによる設計の詳細化、妥当性検証
  - モデルからの自動コード生成による実装
  - テスト・検証におけるモデルの再利用

- 紙の仕様書で不足する情報を補完するために「モデル」を用いることで、
  - 仕様を明確化する
  - 開発プロセス全体のコミュニケーションを改善する
  - 開発の上流工程を重視する

➢ Logical Checkの作成に応用できないか？

参照 ： 柴田 克久 サイバネットシステム, @IT MONOist
http://monoist.atmarkit.co.jp/mn/articles/0903/27/news109.html

実現可能性を探る

- モデルベース開発の特徴より
  - モデルによる仕様の表現・定義「実行可能な仕様書」
    - 実行可能な仕様書を作成すれば出来る
  - モデルのシミュレーションによる設計の詳細化、妥当性検証
    - シミュレーションの定義を状況の再現性ととらえる。SASは再現性のためのソリューションとしては高度なレベルにある。
  - モデルからの自動コード生成による実装
    - マクロ関連の書籍に答えがありそう・・・
  - テスト・検証におけるモデルの再利用
    - うまく開発して再利用すれば工数大幅削減！

➡ **実現可能性は十分にある！**

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

チェックを抽象化してモデルをつくる

抽象化なし	• visit2<visit4-74日 or visit4-46日<visit2 • if EXIT=あり 追跡調査のvisit<中止日 • 「はい」が選択されている場合に、中止の有無で「あり」が選択されていない
第一抽象化	• Visitのallowanceの調査 • チェックボックスと日付の矛盾の調査 • 選択肢間の矛盾の調査
第二抽象化	• 変数間の差と基準値(定数)の比較 • 条件付きで変数間の値を比較 • 変数間の値を比較

Libraryに格納されるモデルの紹介

Generator Model : 001

If DSN1.Var1 Operator Const1 Then RESULT = 1

Generator Model : 002

If DSN1.Var1 – DSN2.Var2 Operator Const1

Then RESULT = 1

Generator Model : 003

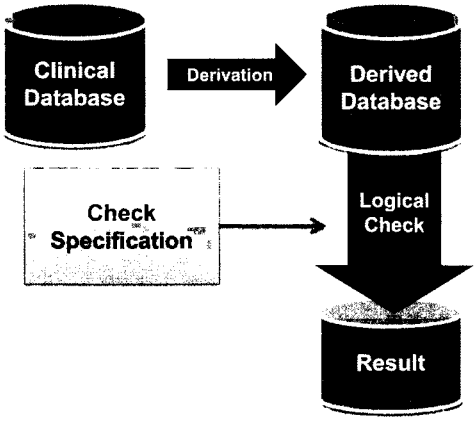
If DSN3.Var3 = Const2 Then

If DSN1.Var1 Operator Const1 Then RESULT = 1

Model based LibraryによるLogical Checkの自動生成と
チェック仕様書作成業務の効率化

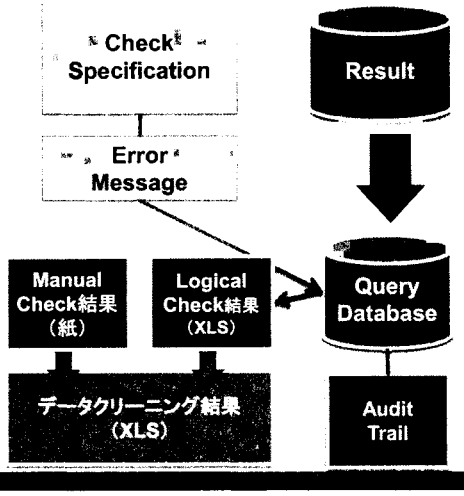
TECHNOLOGY SIDE

Logical Check の構造①



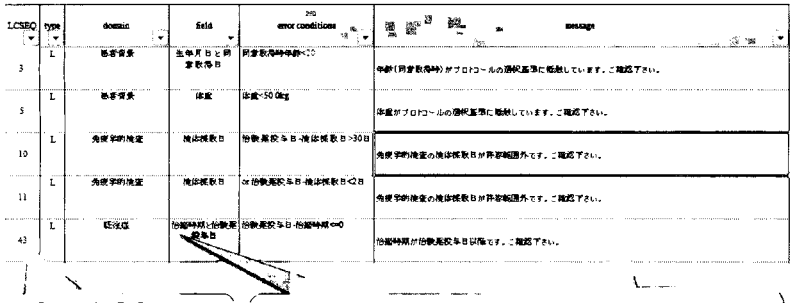
1. Clinical DBに対してチェック用の変数を導出（手動）
2. Logical Check Programは Check仕様書から自分がチェックすべきデータセット名、変数名、基準値などを読み込む
3. 読み込んだ値に対してチェックプログラムを生成。
4. Call Executeで実行する
5. 結果をResultへ格納する

Logical Check の構造 ②



1. Check Spec.からエラーメッセージを読み込み、フォーマットとして登録。
2. Result に対してエラーが検出されたものをquery候補として別のDSへ移行する。
3. Query DSに対してフォーマットを適応し、Logical Checkの結果としてエクセルへ出力

Check仕様書の構造① DM - CRA



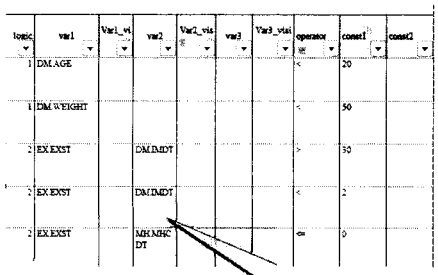
LCSEQ	type	domain	field	error conditions	message
3	L	患者背景	生年月日と同意取得日	同意取得年齢<20	年齢（同意取得時）がプロトコールの適格基準に抵触しています。ご確認下さい。
5	L	患者背景	体重	体重<50.0kg	体重がプロトコールの適格基準に抵触しています。ご確認下さい。
10	L	角度学的検査	検出視取B	治験薬投与B-検出視取B>30日	角度学的検査の検出視取Bが許容範囲外です。ご確認下さい。
11	L	角度学的検査	検出視取B	or 治験薬投与B-検出視取B<2日	角度学的検査の検出視取Bが許容範囲外です。ご確認下さい。
43	L	既往症	治験時期と治験薬投与B	治験薬投与B-治験時期<=0	治験時期が治験薬投与B以降です。ご確認下さい。

CheckのSeq. Number

CRFのページ CRF上の項目

エラーメッセージ

Check仕様書の構造② DM - STAT



logic	var1	Var1_vi	var2	Var2_visi	var3	Var3_visi	operator	const1	const2
1	DM.AGE						<	20	
1	DM.WEIGHT						<	50	
2	EX.EXST		DM.IMDT				>	30	
2	EX.EXST		DM.IMDT				<	2	
2	EX.EXST		MH.MHC DT				ne	0	

Checkの Generate Modelを指定

ドメイン名.変数名の形式で指定

ドメイン名の指定範囲:
SDTM CTのDOMAIN名で標準化（拡張あり）

変数名の指定範囲:
横型のデータセットに対応
命名方法は出来る限り CDISC の命名法に従う

Operator:
全角使用不可。
SASが認識できるもの、なおかつ人も認識しやすいものに統一

基本的な構造

logic	LCSEQ	dsn1	svar1	Var1_visit	operator	const1
1	3	DM	AGE		<	20
1	4	DM	AGE		>=	45
1	5	DM	WEIGHT		<	50
1	6	DM	WEIGHT		>=	85
1	7	DM	BMI		<	18.5
1	8	DM	BMI		>=	25
1	9	DM	FAMILY		=	1

```
data _null_;
  set plan;
  statement =
  'Data temp_result;
   set '||dsn1||';
   LCSEQ = '||compress(""||LCSEQ||"")||' ;
   if '||svar1||' '||operator||' '||const1||' then result = 1;
   else result = 0;
   run;';
 call execute(statement);
run;
```

DSN : Plan
Check仕様書を読み込み、変数とData Set Nameを分離。Model 001の読み込み部分の実行結果。

Visitを考えない場合の生成文。Planの変数を読み込み、statement変数を作成。Call executeでstatement文を実行する。

条件分岐へ拡張

logic	LCSEQ	dsn1	svar1	Var1_visit	operator	const1
1	3	DM	AGE	1	<	20
1	4	DM	AGE	1	>=	45
1	5	DM	WEIGHT		<	50
1	6	DM	WEIGHT		>=	85
1	7	DM	BMI		<	18.5
1	8	DM	BMI		>=	25
1	9	DM	FAMILY		=	1

```
str1 =
  'Data temp_result;
    set '||dsn1||';';
if Var1_visit NE null then str2 =
  'where VISIT = '||Var1_visit||';';
str3 =
  '  LCSEQ = '||compress(""||LCSEQ||"")||' ;
   if '||svar1||' '||operator||' '||const1||' then result = 1;
   else result = 0;
   run;';
  statement = cat(str1, str2, str3);
 call execute(statement);
run;
```

Visit指定が入っているケースを考慮する。

Visitを考える場合は、Visitの指定の有無で生成文を変更する。今回のケースでは、where句のみで十分な制御が可能。Nullの場合は、長い空白が挿入されるが問題はない。

不思議なwarning

文字列が262byteを超えるとwarningが出る。昔の名残らしいがなぜ消去されないのかは謎。

異なるデータセット同士の比較

```
* マクロ変数をsymgetで取得 *;
 mvarc = symget(compress("svar3_"||_n_));
 mvarn = input(mvarc,8.);

* モデル式の中で使用する *;
 if '||svar1||' '||operator||' mvarn then result = 1;
```

```
* 結合する*;
Proc sql noprint;
Create table temp_result as
  select a.SUBJID, ...., b.AESEQ, b.AEPTCD
    from DM a left join AE b
      on a.SUBJID = b.SUBJID;

*構造がいつも同じであればよいのだが・・・。*;
```

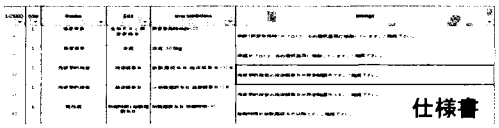
1. マクロ変数を利用する

マクロ変数テーブルを一時的な変数格納庫として利用する。変数はsymgetなどで取得。
- 制御文が変数生成と値の比較の二系統に分かれるので、生成順序などが心配。

2. 結合を利用する

比較する二つのデータセットを結合し、比較する。
- 安定しているが、二つのデータセットの構造の関係や結合キーが安定して存在するかどうかが微妙。

エラーメッセージの出力



仕様書

```
proc sql;
  insert into Query

   select LCSEQ, SUBJID, STAGE, VISIT, TRISEQ,
RESULT, LCSEQ as DOMAIN, LCSEQC as
ERRMSG,  LCSEQ as FLDCD
   from Result
   where RESULT = 1;
quit;
```

Check仕様書をProc formatでエラーメッセージ、CRF項目などをフォーマットとして保存。

フォーマットは、LCSEQをキーにしておき、クエリデータセットにかぶせておく

その後エクセルへ出力。

結果

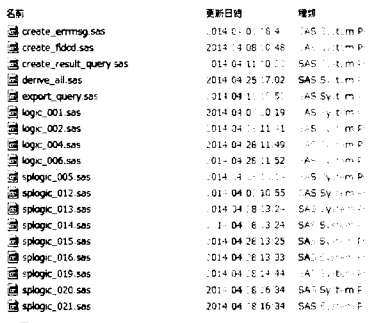
- システムとして構築することができた
  - 再利用するmodelはCSVを実施し、Libraryへと登録して再利用可能としておく。
  - 再利用しないmodelは簡易CSVを実施して利用。
- エラーメッセージ、CRFページ表示、出力、監査証跡機能などを実装
- 拡張するときはmodelを追加する。
  - 次期開発候補はCDISCの縦型データセットに対応できるmodelなど

Model based LibraryによるLogical Checkの自動生成と
チェック仕様書作成業務の効率化

DEMONSTARTION

実施例

- 臨床試験
- 症例数 38
- CRF：紙
- Check数：2173
  （うちSASで処理するもの　：2010）
- DM担当者：新人（DM未経験）



名前	更新日時	種類
create_errmsg.sas	.014 0 0 16 4	SAS
create_fldcd.sas	2014 4 08 0 48	SAS
create_result_query.sas	014 04 11 10 0	SAS
derrve_all.sas	2014 04 25 17.02	SAS
export_query.sas	.014 04 1 5	SAS
logic_001.sas	2014 04 0 0 19	SAS
logic_002.sas	.014 04 1 11 1	SAS
logic_004.sas	2014 04 26 11 49	SAS
logic_006.sas	014 04 26 11 52	SAS
splogic_005.sas	.014 4	SAS
splogic_012.sas	.01 04 0 10 55	SAS
splogic_013.sas	.014 04 8 3 2	SAS
splogic_014.sas	04 6 3 24	SAS
splogic_015.sas	014 04 26 13 25	SAS
splogic_016.sas	014 04 8 13 33	SAS
splogic_019.sas	.014 04 8 14 44	SAS
splogic_020.sas	201 04 8 6 34	SAS
splogic_021.sas	2014 04 8 16 34	SAS

Libraryからコピーされた**Logical Check Program**。今後はバッチファイル処理などで効率的に実施出来るようにする予定。

➡ **結果として18のモデルで全てのCheckを生成！**

考察 －工数変化とその他の変化-

- DM担当者にとっては無事にCheck仕様書が作成できた。Checkの有り方はモデルが示すため作業自体は効果的に進んだ。ただモデルを定める作業が入ってしまったため、若干工数がかさんだ。しかしモデルが安定すればもう少し工数が減ると予想。
- STATの工数はProgram Libraryに新規modelを追加するときのみ増大するため、相当の工数が減少する予定。
- 部内のやり取りはモデルベースで行うことができるため、仕様書における変数や基準値の話題がメインに。 → ！？

考察 ② - モデルが促進するコミュニケーション -

- 従来のSASプログラミング工程にmodel basedの概念を持ち込んで開発した結果、DMおよびSTAT担当者の物の見る目が変化した。
  - 前:どういうCheckを作ればよいか
  - 後:どういうモデル、変数を作れば、使えばよいか
  - ➢ いつの間にか変化が起こり、モデルを中心としたコミュニケーションが実施されていた
  - ➢ 結果としてモデルは明確な判断基準と齟齬のより少ないコミュニケーションをDM・STATグループへ提供した

結論 - コンストラクタル法則の観点より -

- モデルは
  - ➢結局のところ高次元言語のように振る舞った
- モデルが加速させた業務プロセス
  - ➢情報伝達プロセス
- Model based開発を組み込んだこと
  - ➢コンストラクタル法則が示す進化の方向と同等

➡おそらく多くの場面で応用が可能と考えられる

Reference

- Adrian, Bejan, *『流れとかたち— 万物のデザインを決める新たな物理法則』* 柴田裕之 訳
- Art, Carpenters, *『Carpenter's Complete Guide to the Sas Macro Language』*
- Robert, Virgile, *『SAS Macro Language Magic: Discovering Advanced Techniques』*
- SAS 9.3 *SQLプロシジャ ユーザーガイド*
- SAS 9.3 *マクロ言語: リファレンス*

ご静聴ありがとうございました！

And special thanks to Adrian Bejan !

PMDAへの承認申請時 CDISC標準電子データ提出に向けた 社内標準のリモデリング

神谷 亜香里, 坂井 絵理, 惟高 裕一,
北西 由武, 角谷 伸一, 小坂 明子
塩野義製薬株式会社 解析センター

Remodeling Shionogi standard for clinical data to meet the requirement of PMDA based on CDISC standard

Akari Kamitani, Eri Sakai, Yuichi Koretaka, Yoshitake Kitanishi, Shinichi Kakutani, Akiko Kozaka
Biostatistics Department, SHIONOGI&CO.,LTD.

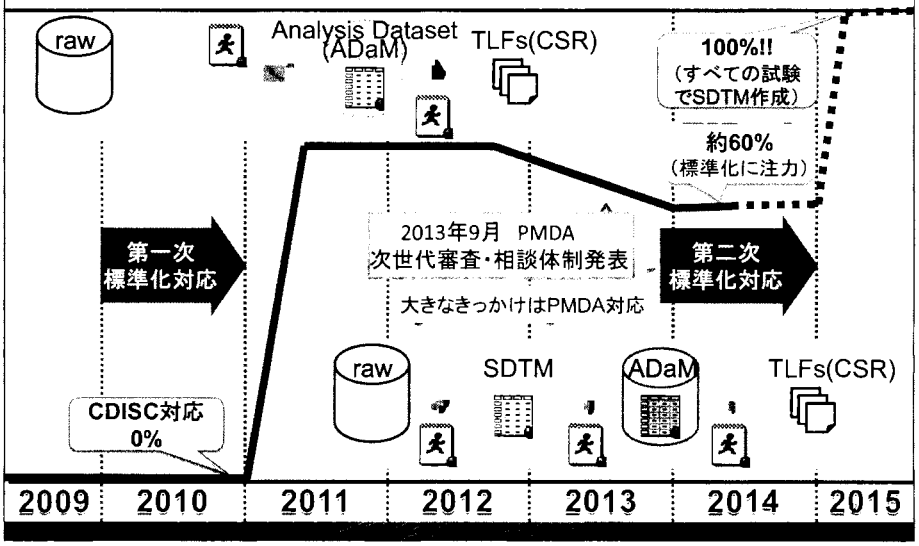
医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

PMDAへの承認申請時電子データ提出に伴い,
Raw data→SDTM→ADaM→TLFの流れで効率的に
業務を行うために社内標準の手順を再構築中である.
本発表では現行手順と検討ポイントを紹介する.

キーワード: CDISC, SDTM, ADaM, PMDA, FDA,
社内標準, 効率化, CDI, SASマクロ

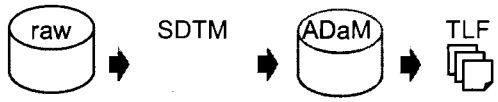
シオノギでの解析手順の変遷のイメージ



医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

本日の内容

☆シオノギのこれまでの取り組み



☺ 第一次CDISC標準化対応 For FDA(2010年)

➢体制・作成方法の「**確立**」

✏第二次CDISC標準化対応 For All Over the World(2013年～)

➢体制・作成方法の「**リモデリング**」

☆日本独特の問題

☆悩みの共有

4

第一次CDISC標準化対応 For FDA (2010年)

raw SDTM ADaM TLF の確立

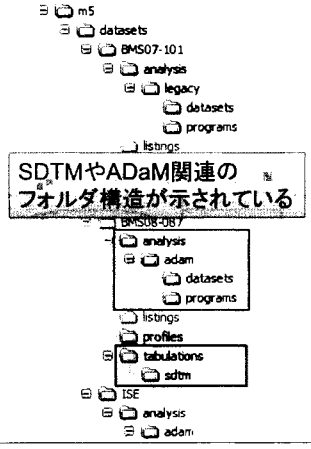
☆要点とシオノギでの工夫

- ➢体制の確立を最優先！
  - ✓IGの理解
  - ✓フォルダ構造の工夫
  - ✓SDTMの工夫
    （SASプログラミングスキルのない人でも作成可能に！）
  - ✓ADaMの工夫

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

フォルダ構造の工夫 ～eCTDを意識～

FDAがeCTDで求めるmodule5の構造



シオノギが作ったフォルダ構造

- FDAが提示した構造に加え，TLFやSAP，TLF shellsを格納するフォルダを追加した社内独自のフォルダ構造を構築
  ➔フォルダをtemplate化
  (全115folders)
- 上記で構築したフォルダ別，役割別にアクセス権を細かく定義

Source: FDA. Study Data Technical Conformance Guide.; February 2014.

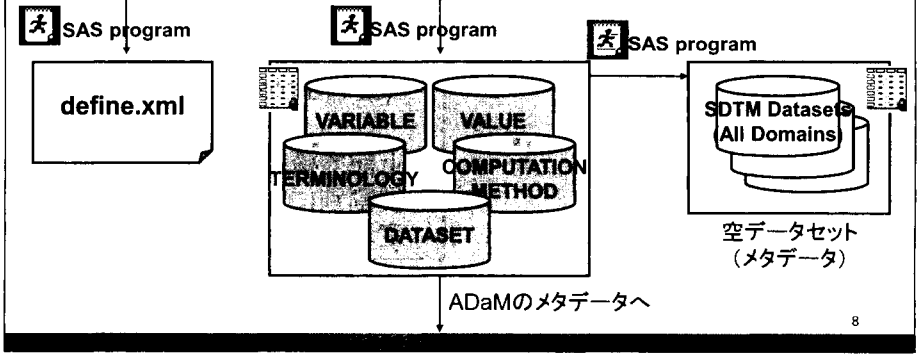
6

SDTM作成・QCの工夫

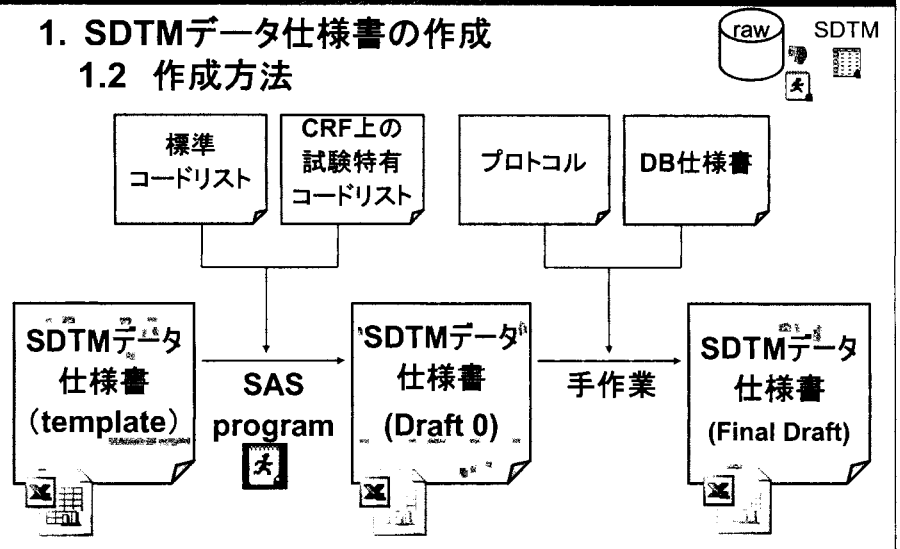
1. SDTMデータ仕様書の作成
   - 1.1 3つの用途
   - 1.2 作成方法
2. SDTMの作成
   - 2.1 SDTM作成用SASマクロ・関数
   - 2.2 CDIとSDD
3. SDTM仕様書とSDTMのQC
   - 3.1 チェックリストに基づく目視チェック
   - 3.2 SASプログラムによるチェック
   - 3.3 OpenCDISC
   - 3.4 DM担当者による目視チェック

1. SDTMデータ仕様書の作成
1.1 3つの用途

1. define.xmlの作成
2. SDTM作成をサポートするメタデータの作成
3. 空のSDTMデータセットの作成
   （CDIでのアウトプットデータ）

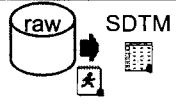


1. SDTMデータ仕様書の作成
1.2 作成方法



医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

2. SDTMの作成
2.1 SDTM作成用SASマクロ・関数



[マクロ]

- *decode*：メタデータTERMINOLOGYを用いて，指定したコードリストのコード値（CODE）に対応したCODEDVALUE列のデータ値を格納する.
- *sequence_generator*：指定変数（主に--SEQ）への連番の格納に用いる.

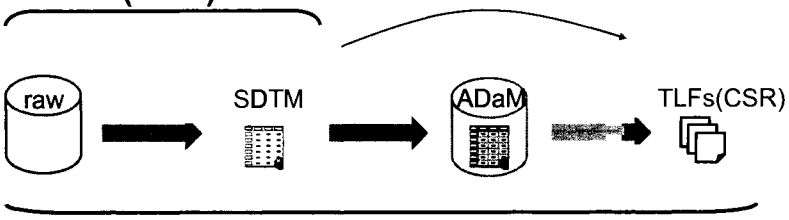
[関数]

- *to_iso*：ISO8601形式の日付を，年，月，日それぞれの数字変数から作成する.
- *diff_day*：ISO8601形式の日付同士の日数差を計算する.

10

2. SDTMの作成
2.2 CDIとSDD ~位置づけ~

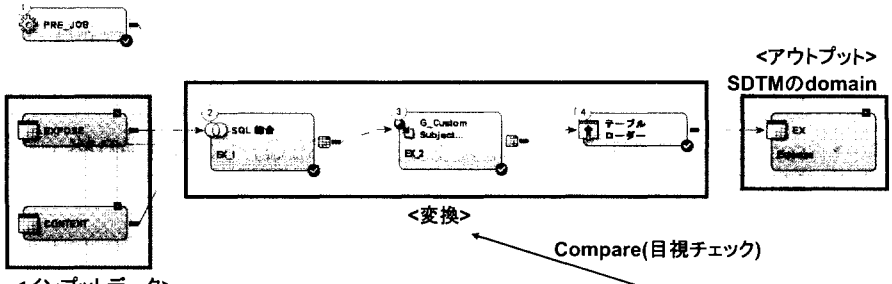
Clinical Data Integration
(V2.3)



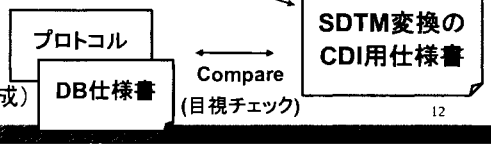
SAS Drug Development
(V3.5)

2. SDTMの作成
2.2 CDIとSDD ~CDIによるSDTM作成~

・GUIベースで，SASプログラミングスキルがなくても，作成可能！

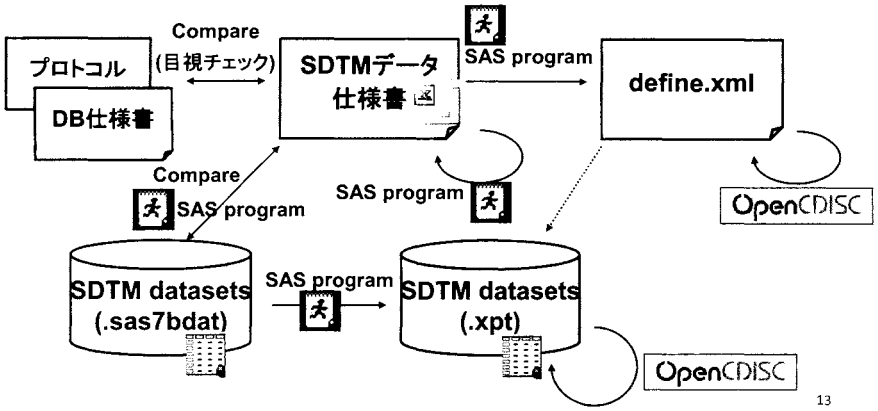


・シングルプログラムで可能
（SDTM変換のCDI用仕様書を作成）



3. SDTM仕様書とSDTMのQC

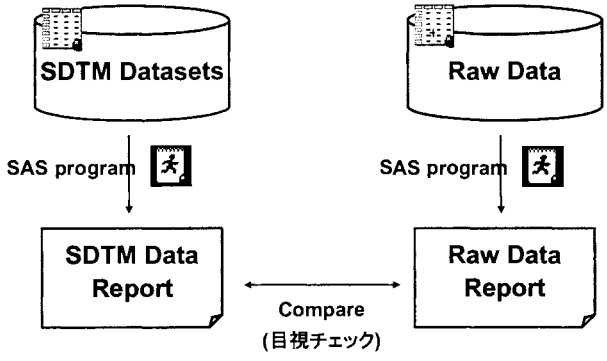
3.1 チェックリストに基づく目視チェック
3.2 SASプログラムによるチェック
3.3 OpenCDISC



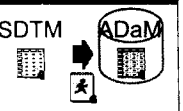
医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

3. SDTM仕様書とSDTMのQC

3.4 DM担当者による目視チェック



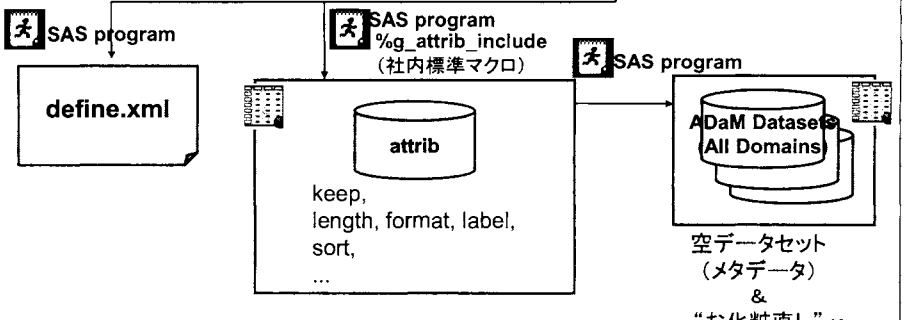
ADaM作成・QCの工夫



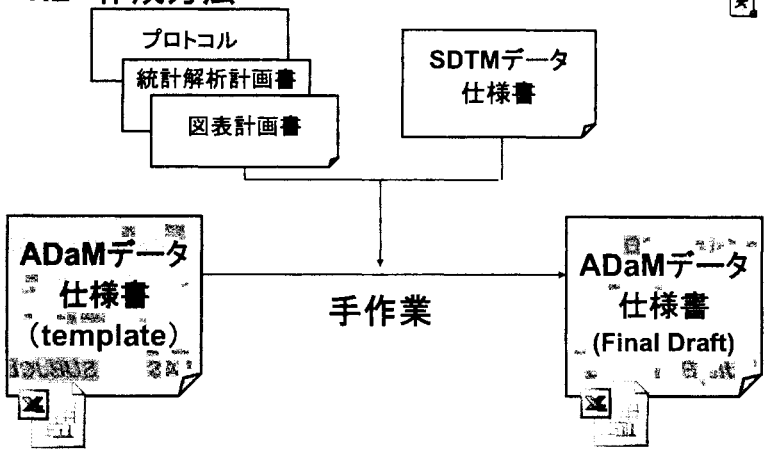
1. ADaMデータ仕様書の作成
   1.1 3つの用途
   1.2 作成方法
2. ADaMの作成
   2.1 ADaM作成用SASマクロ
3. ADaM仕様書とADaMのQC
   3.1 チェックリストに基づく目視チェック
   3.2 SASプログラムによるチェック

1. ADaMデータ仕様書の作成
1.1 3つの用途

1. define.xmlの作成
2. ADaM作成をサポートするメタデータの作成
3. 空のADaMデータセットの作成



1. ADaMデータ仕様書の作成
1.2 作成方法

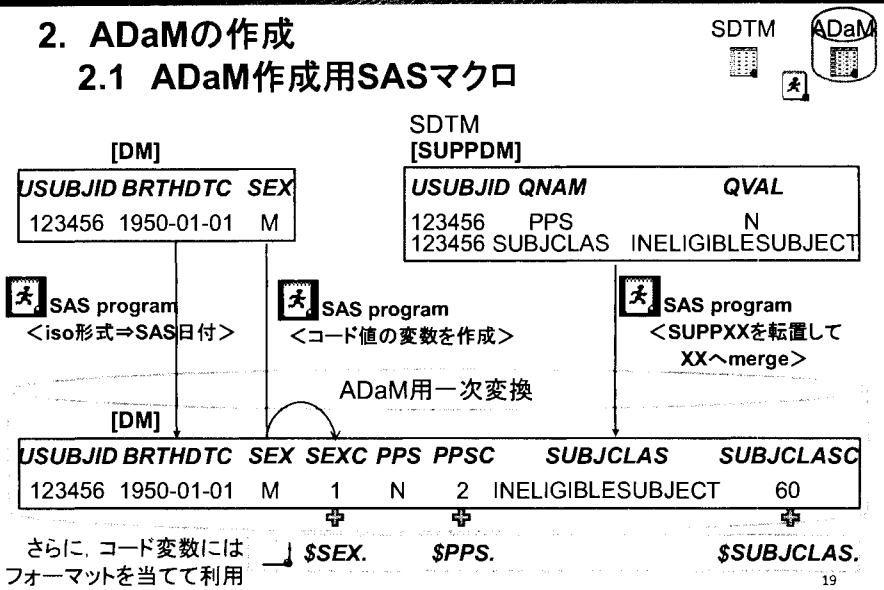


2. ADaMの作成
2.1 ADaM作成用SASマクロ

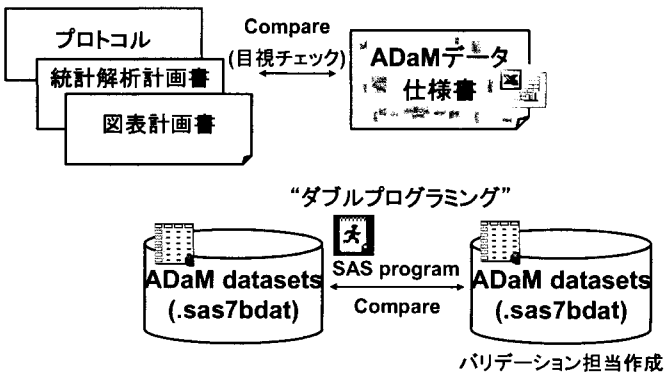
[マクロ]

- *sdtm_supp_merge*:SDTM SUPPドメインを親ドメインにマージしたSASデータセット作成
- *sdtm_encode_sub*:テキスト値をコード値に変換
- *is8601dt*:<SDTMデータ>内のISO8601形式で作成された全ての日付データ(文字型)をSAS日付値またはSAS日時値データに変換

2. ADaMの作成
2.1 ADaM作成用SASマクロ



3. ADaM仕様書とADaMのQC
3.1 チェックリストに基づく目視チェック
~統計解析担当者，プログラム担当者，
バリデーション担当者によるチェック~
3.2 SASプログラムによるチェック



第二次CDISC標準化対応
For All Over the World (2013年～)

すべての試験を raw → SDTM → ADaM → TLF で

実施/実現するための体制をリモデリング

☆第二次標準化を行わなければならない課題および解決

- ✓2013年9月 PMDAの発表「次世代審査・相談体制」
- ✓SDTMを介さずにADaMを作成する場合の問題
- ✓SDTM作成担当者の不足

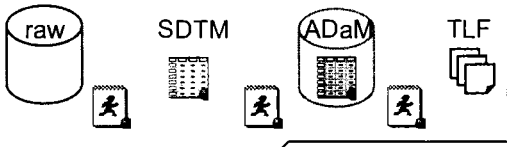
⇩

- ✓手順の一本化
- ✓作成の工夫
- ✓リソースの補強

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

正道か？オプションか？「再考」

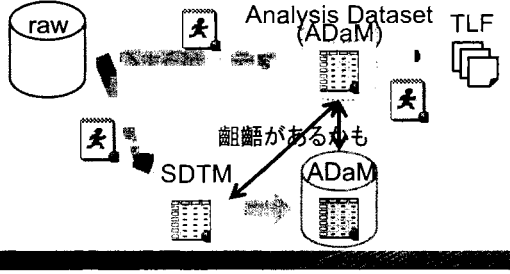
①正道



- ✓ CSR(TLF)からSDTMへのトレーサビリティを保つことが重要.
- ✓ SDTMは,
  - とくに安全性の統合解析で役立つ.
  - 海外とデータをやり取りしやすい.
  - 申請時絶対に不要な試験はない.

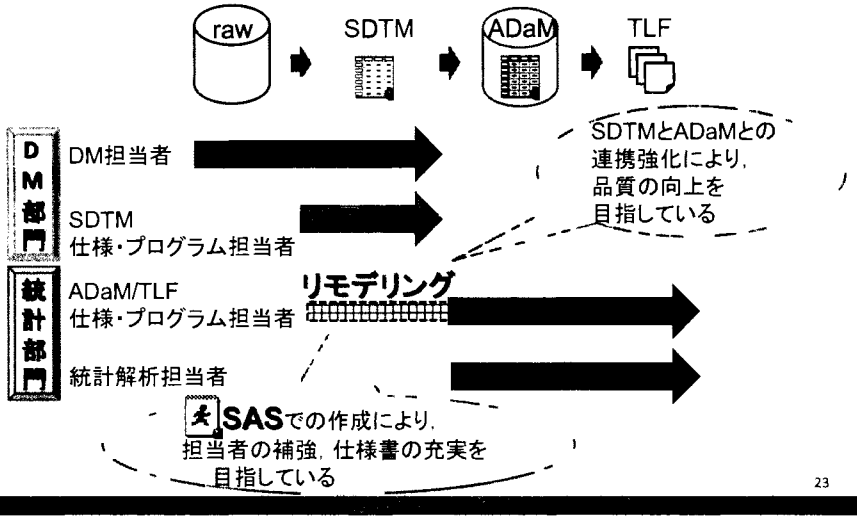
手順は一本のほうがよい

②オプション



- ✓ 必要なときにSDTMを作成すればよいので, CSR作成時に常にタスクをかけなくてよい.

体制と作成方法のリモデリング



医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

日本独特の問題

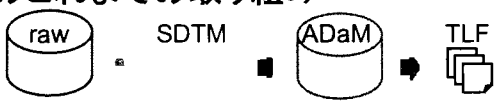
- SDTM
  - CRFや日誌の質問および回答が日本語の場合，「オリジナル」は日本語と考えるのでは？
  - 薬剤コーディング辞書：医療用医薬品名データファイルとWHO-DDで格納方法が異なる
- ADaM
  - 帳票を日本語で作成する場合，SASのフォーマットで対応
- 共通
  - 日本語などの2バイト文字を利用していると，システムがうまく動かないものがある（JMP Clinicalもその一つ）．
  - ℃やμなどの対応は？
  - 2バイト文字を受け入れ可能なシステムであっても，SJISとUTF-8など文字コードが違っていれば，文字化けの可能性がある．
    「’s」が，「痴」に化ける！

その他の悩み

- 局所最適化 ⇒ 全体最適化
  - 解析センター外とのコラボレーションが必要
- バージョン対応とその管理と更新
  - SDTMのIG
  - Controlled Terminology
- CDISC標準の社内教育
  - 解析センター内(DM部門, 統計解析部門)
  - 解析センター外
- PMDAの動向のWatch

まとめ

☆シオノギのこれまでの取り組み



☺ 第一次CDISC標準化対応 For FDA(2010年)
- ✓体制の確立を最優先(IGの理解, CDIの利用)
- ✓工夫点の紹介(テンプレート, マクロ, フォルダ構造)

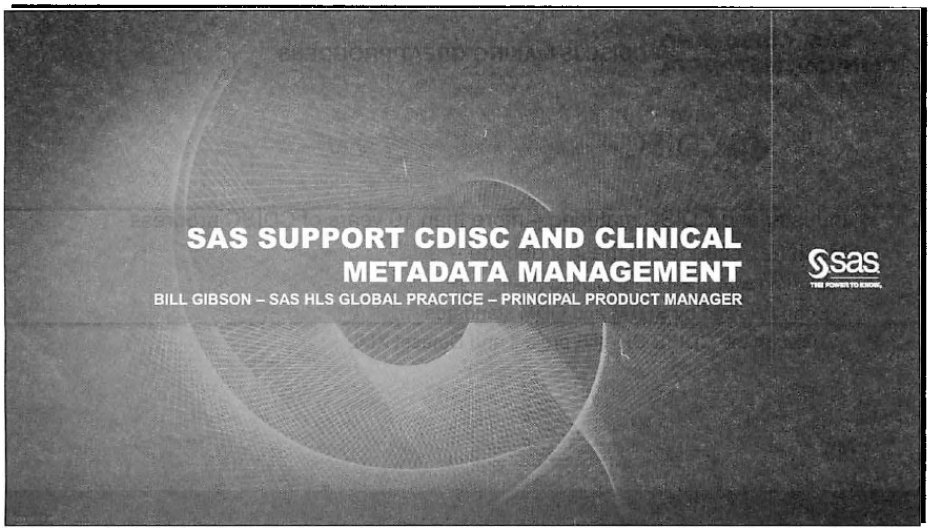
✏第二次CDISC標準化対応 For All Over the World(2013年～)
- ✓2013年9月PMDAの「次世代審査・相談体制」発表をきっかけ
- ✓すべての試験でSDTMを作成して, 一本の手順で業務を行うため, 体制をリモデリング

☆日本独特の問題

☆悩みの共有

参考資料

1. FDA. Study Data Technical Conformance Guide.; February 2014.
   Available from: URL : http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf
2. PMDA. PMDAの次世代審査・相談体制に関する説明会.; September 2013.
   Available from: URL : http://www.pmda.go.jp/operations/shonin/info/iyaku/jisedai/file/20130910-pmda-jisedai.pdf
3. 北原孝志, 東島正堅, 北西由武, 吉田祐樹. 解析業務プロセスにおいて効率的な仕様書作成とDefine.xmlへの変換. SASユーザー総会. 兵庫. 2011.
4. 惟高裕一, 藤原正和, 北西由武, 吉田祐樹. SASを使った情報管理事例 ～そしてリスクをやっつけろ～. SASユーザー総会. 東京. 2013.
5. 豊泉樹一郎, 北西由武, 吉田祐樹, 平井健太. FDA Submissionのためのdefine.pdf作成事例-SAS®によるファイル変換のAutomation化-. SASユーザー総会. 東京. 2013.
6. 渡邉慶. SASソリューションを利用した臨床試験データリポジトリの構築. SAS Life Science Forum. 東京. 2011.
7. Holland C, Shostak J. Implementing CDISC Using SAS®: An End-to-End Guide. SAS Institute INC (NC); 2012.



SAS, CDISC, AND CLINICAL METADATA OVERVIEW AND INTRODUCTION

Abstract:
SAS's current and future support of CDISC, such as possible integration with CDISC SHARE, and enhanced metadata management in clinical data repository based on customers' use cases and product road maps.

Main Items:

- Industry and CDISC maturing – more than 10 years of CDISC progress
- SAS support of CDISC
- Standards updated frequently – updates without new SAS versions
- Roadmap (recent, next release, next 3 years):

sas THE POWER TO KNOW

SAS, CDISC, AND CLINICAL METADATA

CDISC IS MAKING GREAT PROGRESS

- Industry and CDISC maturing – more than 10 years of CDISC progress
- Much has changed in the last decade –
  - Starting with creation of CDISC itself
  - Soon followed by SDTM and ODM standards
  - Through the CDISC SHARE project and Dataset XML

SAS, CDISC, AND CLINICAL METADATA

SAS SUPPORTS CDISC AND OUR CUSTOMERS

- SAS provides formal support for many CDISC standards:
- SDTM versions 3.1.1, 3.1.2, 3.1.3 and 3.2
- ADaM 2.1
- SEND 3.0
- define.xml (CRT-DDS 1.0 and Define 2.0)
- ODM 3.1 and 3.1.1
- Supported in SAS Drug Development and SAS Clinical Data Integration via SAS Clinical Standards Toolkit

- Dataset XML, CDASH, and SHARE support coming soon

SAS, CDISC, AND CLINICAL METADATA — HOW TO MANAGE FREQUENT UPDATES

- Standards updated frequently – updates without new SAS versions
- Loading ODM terms first supported in Clinical Standards Toolkit 1.5
- Download desired terminology list from CDISC/NCI in CDISC ODM format
- Use process defined in Lex Jansen's PharmaSUG 2013 paper HT06-SAS

- CDISC SHARE will provide data standards structures and other metadata
- Once available, CDISC SHARE extracts will be importable in a similar way
- CDISC SHARE to provide define.xml and ODM formats which can be used now

sas

CLINICAL DEVELOPMENT ROADMAP

RECENT RELEASES, NEXT RELEASE, NEXT 3 YEARS

sas

SAS, CDISC, AND CLINICAL METADATA — RECENT RELEASES

- Clinical Data Integration 2.5 released in March, 2015
- Clinical Standards Toolkit 1.6 released in Feb, 2015
- Key new features
  - SDTM 3.2 Support
  - Define 2.0 Support
  - Version 5 Transport File Handling
  - Creating Metadata from Define-XML Files
  - Incremental Update of Data Standard Metadata from CST

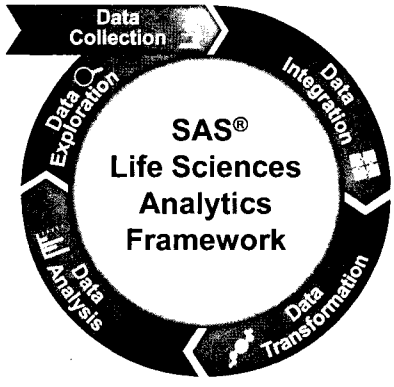
SAS, CDISC, AND CLINICAL METADATA — NEXT RELEASE

- Clinical Data Integration 2.6 and Clinical Standards Toolkit 1.7
- Planned release 2015 Q1
- Completion of define metadata (1.0 and 2.0)
- Support for CDISC Dataset XML
  - SAS participating in FDA pilot for Dataset XML (replacement for SAS V5 Transport)
- Incorporation of CDISC CDASH data collection standard

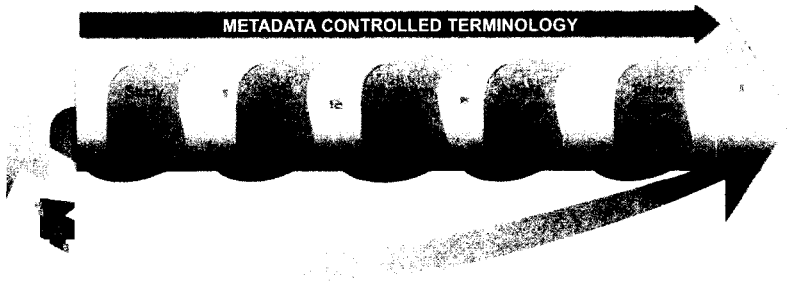
SAS, CDISC, AND CLINICAL METADATA NEXT 3 YEARS (OR BEYOND NEXT RELEASE)

- Releasing new SAS offering 2015, Q2 – Life Science Analytics Framework
- Combined feature set with new metadata management capabilities
  - Swith CDASH-SDTM/SEND-ADaM definitions to SAS dataset
  - Ability to create define.xml (10/2.0) for appropriate standards
  - All versioned and controlled
  - Modifications made via rule-based wizards (like in CDI today)
  - Clear understanding of study-study linkages/portfolio of studies a core concept

SAS, CDISC, AND CLINICAL METADATA END-TO-END ANALYTICS

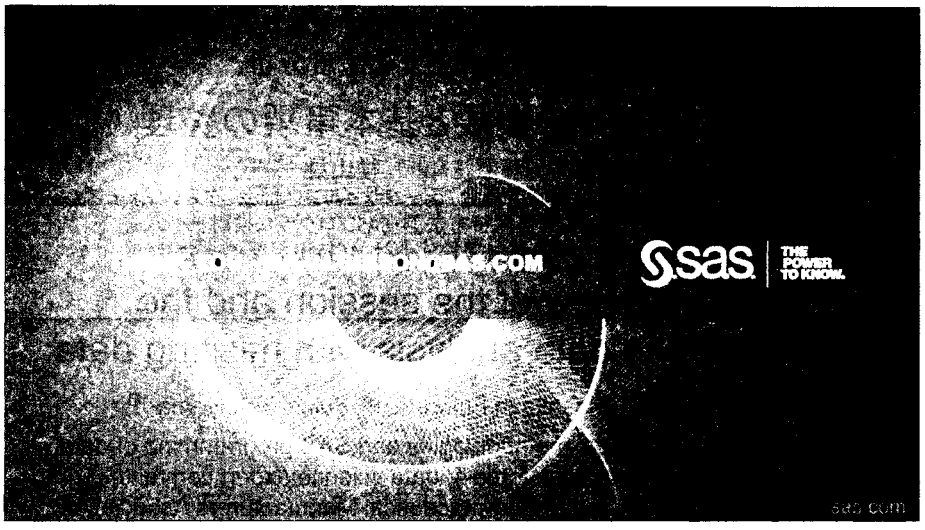


EXAMPLE: DATA FLOW AUTOMATED FROM STUDY METADATA



SAS, CDISC, AND CLINICAL METADATA IMPORTANT ITEMS COVERED

- CDISC maturing – existing standards changing slowly but new ones added
- SAS supports CDISC and customers using their standards
- Update standards without installing new SAS versions
- SAS plans many new capabilities for future software releases


sas
THE POWER TO KNOW.
sas.com

【企画セッション】欠測のあるデータに対する各種解析手法と
欠測メカニズムに対する感度分析

(1)セッションの概要と基本事項の整理

土居正明[1)2)] 藤原正和[1)3)] 横山雄一[1)4)]
1)日本製薬工業協会 医薬品評価委員会 データサイエンス部会 タスクフォース4
欠測のあるデータに対する解析方法論・SASプログラム検討チーム
2) 東レ株式会社 3) 塩野義製薬株式会社 4) 持田製薬株式会社

The overview of the session and the introduction to data analysis with missing data.

Masaaki Doi[1)2)], Masakazu Fujiwara[1)3)], Yuichi Yokoyama[1)4)]
1) The team for statistical methodologies and SAS programming of data analysis with missing data, task force 4, data science expert committee, drug evaluation committee, Japan Pharmaceutical Manufacturers Association.
2) Toray industries, Inc. 3) Shionogi & Co., Ltd.
4) Mochida Pharmaceutical Co., Ltd.

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

欠測のあるデータの解析の基本的な用語を整理しつつ，企画セッション全体の概説を行う．
特に，欠測メカニズム，LOCFの使用に対する検討，MMRMの導入，感度分析の大まかな枠組み等について述べる．

キーワード:欠測メカニズム，感度分析，LOCF，MMRM，選択モデル，パターン混合モデル，共有パラメータモデル

発表内容

1. セッションの概要
2. 記号の整理と状況設定
3. 欠測メカニズムの解説
4. 解析に用いるデータ・プログラムの紹介
5. 主解析について(LOCF・MMRM・その他)
6. 感度分析の概要
7. まとめ

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

1. セッションの概要

セッションの概要：背景

欠測のあるデータの解析の現状とよくある疑問
（経時データを扱う臨床試験を想定）

①LOCFは使ってはダメ？
- → どういう状況でも使えない？
- → 主解析としてはダメ？

②主解析にMMRMも使われるようになってきた.
- → MMRM の厳密な定義は？
- → MAR であることの妥当性は？
- → 他の解析方法は？

③感度分析は何をすればよい？
- → そもそも「何に対する」感度を検討するべき？

④SASで容易に実行可能？
- → 自分で膨大なプログラミングが必要？

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

セッションの概要：本セッションの目的

National Research Council (2010) "The Prevention and Treatment of Missing Data in Clinical Trials." （通称**NASレポート**）

→ **考え方・試験計画・主解析・感度分析に対する提言**

本セッションでは，NRC (2010) をベースに，
- ・欠測のあるデータに関する基本的事項を整理する
- ・主解析となる解析やその他の解析方法を整理する
- ・欠測メカニズムに対する感度分析の方法を紹介する

◎理論の理解とSASでの実行方法をみる.

セッションの概要：セッションの構成

(1) セッションの概要と基本事項の整理（本発表）

(2) 解析手法の解説1（選択モデル・MMRM）

(3) 解析手法の解説2
（パターン混合モデル・共有パラメータモデル）

(4) 欠測メカニズムに対する感度分析

(5) まとめと質疑応答

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

2. 記号の整理と状況設定

略語一覧

（欠測メカニズム）

- MCAR : Missing Completely At Random
- MAR : Missing At Random
- MNAR : Missing Not At Random

（解析方法の名称）

- SM : Selection Model（選択モデル）
- PMM : Pattern Mixture Model（パターン混合モデル）
- SPM : Shared Parameter Model（共有パラメータモデル）
- MMRM : Mixed effect Models for Repeated Measures
- MI : Multiple Imputation
- IPW : Inverse Probability Weighting

注意点：用語・記号が紛らわしい

- MAR：文献ごとに定義が異なる
- 感度分析？感度解析？
- MMRMの厳密な定義は？
- 欠測識別変数は R_{ij}? M_{ij}? 欠測のときに 0? 1?

2. 記号の整理と状況設定

対象となるデータ：経時データ（連続値）

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_i^o \\ \mathbf{Y}_i^m \end{pmatrix} \quad (i = 1, \cdots, N)$$

n_i ：被験者 i の（計画された）測定時点

N ：被験者数

$\mathbf{Y}_i^o$ ：観測データ

$\mathbf{Y}_i^m$ ：欠測データ

欠測識別変数

$$R_{ij} = \begin{cases} 1 & \text{被験者 } i \text{ の } j \text{ 時点でのデータが観測された} \\ 0 & \text{被験者 } i \text{ の } j \text{ 時点でのデータが欠測} \end{cases} \qquad \mathbf{R}_i = \begin{pmatrix} R_{i1} \\ \vdots \\ R_{in_i} \end{pmatrix}$$

$$D_i = \sum_{i=1}^{n_i} R_{ij} + 1 \quad \rightarrow$$

単調な欠測（後で定義）の場合，
被験者 i は時点 D_i で脱落，
完了例は $D_i = n_i + 1$ ，

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

データの種類と尤度

・完全データ (Full data)

$(\mathbf{Y}_i, \mathbf{R}_i) = (\mathbf{Y}_i^o, \mathbf{Y}_i^m, \mathbf{R}_i)$ ◎ $\mathbf{Y}_i^m$は手に入らない

$$\tilde{L}(\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{N} f(\mathbf{Y}_i^o, \mathbf{Y}_i^m, \mathbf{R}_i | \boldsymbol{\theta}, \boldsymbol{\psi})$$

・観測データ (Observed data)

$(\mathbf{Y}_i^o, \mathbf{R}_i)$ ◎手に入る

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{N} \int f(\mathbf{Y}_i^o, \mathbf{Y}_i^m, \mathbf{R}_i | \boldsymbol{\theta}, \boldsymbol{\psi}) d\mathbf{Y}_i^m$$

欠測のパターン

単調な欠測：1回欠測した後は，全て欠測．
→ 中止など
→ **本セッションでは，単調な欠測を仮定**

非単調な欠測：1回欠測した後に，再度観測データあり．
→ ある時点の来院の不備など

○：観測
×：欠測

	時点1	時点2	時点3	時点4	
症例1	○	○	×	×	単調な欠測
症例2	○	○	○	×	
症例3	○	○	×	○	非単調な欠測

3. 欠測メカニズムの解説

欠測メカニズム

扱いやすい

MCAR (Missing Completely At Random)

$$f(\mathbf{R}_i|\mathbf{Y}_i,\boldsymbol{\psi}) = f(\mathbf{R}_i|\boldsymbol{\psi})$$

MAR (Missing At Random)

$$f(\mathbf{R}_i|\mathbf{Y}_i,\boldsymbol{\psi}) = f(\mathbf{R}_i|\mathbf{Y}_i^o,\boldsymbol{\psi})$$

MNAR (Missing Not At Random)

$$f(\mathbf{R}_i|\mathbf{Y}_i,\boldsymbol{\psi}) \neq f(\mathbf{R}_i|\mathbf{Y}_i^o,\boldsymbol{\psi})$$

一般的

欠測メカニズムの例(脱落確率のモデル)

◎2時点(Y_{i1}, Y_{i2}). ベースライン(Y_{i1})は常に観測される.

MCAR $f(R_{i2}=0|Y_{i1},Y_{i2},\psi) = \psi$ ※全て単調な欠測を仮定

MAR $f(R_{i2}=0|Y_{i1},Y_{i2},\psi_0,\psi_1) = \dfrac{\exp(\psi_0+\psi_1 Y_{i1})}{1+\exp(\psi_0+\psi_1 Y_{i1})}$

$R_{i2}=0$ のとき欠測

MNAR

$$f(R_{i2}=0|Y_{i1},Y_{i2},\psi_0,\psi_1,\psi_2) = \frac{\exp(\psi_0+\psi_1 Y_{i1}+\psi_2 \boxed{Y_{i2}})}{1+\exp(\psi_0+\psi_1 Y_{i1}+\psi_2 \boxed{Y_{i2}})}$$

$(\psi_2 \neq 0)$

SASによる欠測のあるデータの発生方法(例)

1. 完全データを発生させる.
2. 欠測識別変数を発生させる.
   ・脱落確率のモデル + Bernoulli乱数
3. 欠測識別変数の値が 0 となったデータを欠測にすることで, 観測データを作成する.

欠測メカニズムの注意点

◎定義が文献によって異なる
- ・共変量を考慮する?
- ・共変量をデザイン変数と補助変数に分ける?
- ・変量効果を考慮する?

・"MAR"という表現では不正確(Seaman, 2013)
- ・realised MAR
- ・everywhere MAR

※本セッションでは扱わない

欠測メカニズムの注意点

◎多くの場合MCARは非現実的

◎MARかMNARは観測データからは区別できない

◎欠測理由の調査が重要

- ・転居による脱落ならMCAR?
- ・原疾患の悪化や有害事象ならMAR, MNAR?
  - → 症例毎にMCAR, MAR, MNARが異なる
  - → 理論的には，MNARの症例が1例でも存在すれば全体としてMNAR

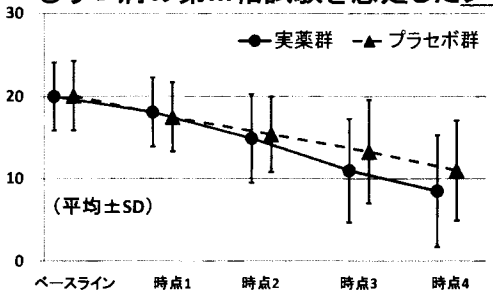
◎試験デザインも重要

- ・原疾患の悪化による中止の場合
  - ・応答変数の測定を頻繁に行う → MAR?
  - ・応答変数の測定が稀 → MNAR?

4. 解析対象データ・プログラムの紹介

解析対象データ

◎うつ病の第III相試験を想定したシミュレーションデータ



- 主要評価項目：HAM-D
  → スコア低下：改善
  （解析には変化量使用）
- 実薬群 vs プラセボ群
  ・1群100例（ベースライン時）

◎単調な欠測のみ

	ベースライン		時点1		時点2		時点3		時点4	
	例数	平均 (SD)	例数	平均 (SD)	例数	平均 (SD)	例数	平均 (SD)	例数	平均 (SD)
実薬群	100	20.0 (4.1)	93	18.1 (4.2)	89	14.9 (5.4)	84	11.0 (6.3)	83	8.5 (6.8)
プラセボ群	100	20.1 (4.2)	90	17.5 (4.2)	87	15.4 (4.6)	85	13.3 (6.3)	80	11.0 (6.1)

21

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

使用するプログラム

MNARの解析には，Missingdata.org.uk の
DIA working group の公開マクロを使用

＜Missingdata.org.uk＞
http://missingdata.org.uk/

＜マクロ＞
http://missingdata.lshtm.ac.uk/index.php?view=category&id=61%3Amnar-methods&option=com_content&Itemid=137

＜窓口＞
Jonathan Bartlett, James Carpenter, Mike Kenward

22

5. 主解析について

LOCFに対する検討

- 単一の値による補完方法は欠測データに対する主要なアプローチとして現在でも広く使用されている
  - 実行が容易
- NRC (2010)とEMAガイドライン(2010)で共通して指摘されている問題点
  - 中止後に値が変わらないという**強い仮定**を置いており，その仮定が妥当でない場合はバイアスが生じる.
    - MCAR, MAR, 及びMNARの条件だけで，LOCFの適合の可否は判断すべきでない.
  - バイアスの方向は**必ずしも保守的とは限らない**.
  - 単一の値による補完であるため，不確実性を無視することで標準誤差が過小評価される可能性がある.

LOCFに対する検討

- **LOCF使用上の注意**
  - NRC (2010)
    - 仮定が正当化できない限り，主要解析として使用すべきではない．
  - EMAガイドライン(2010)
    - 明らかに保守的な場合，説得力のあるエビデンスとなり得る．
- 日本においては，EMAガイドラインのように欠測データの取り扱いに関する規制当局の考え方を示したガイドラインは存在しない．

25

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

主解析の全般的な傾向

・欠測メカニズム
- MAR を仮定することが多い．
  - MARを仮定することは正当化できるか？

・解析手法
- MMRMを使うことが増えてきた．
  - ・MMRMの厳密な定義は？
- 他の解析方法は？

→ (2) 解析手法の解説1（SM・MMRM）
(3) 解析手法の解説2（PMM・SPM）

26

MMRM (Mixed effect Models for Repeated Measures)

$$\mathbf{Y}_i \sim \mathrm{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}), \quad \mathbf{V} = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^{\mathrm{T}} + \Sigma_i$$

・厳密にどういうモデル？
  ・共変量にベースラインは入れる？
    ベースラインと時点の交互作用は？
  ・相関構造は？
  ・変量効果は？
  ・"MMRM"という言葉を使うことに対する批判も
    (Mallinckrodt, 2013).

→ (2) 解析手法の解説1（SM・MMRM）

27

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

MMRM以外の解析方法は？

◎ **MMRMは色々仮定したもとで妥当**
  ・応答変数の正規性
  ・欠測がMAR

→ **他のモデル・解析方法は？**

◎選択モデル (Selection Model, SM)
◎ Multiple Imputation (MI)
◎パターン混合モデル(Pattern-Mixture Model, PMM)
◎共有パラメータモデル(Shared Parameter Model, SPM)

→ (2) 解析手法の解説1（SM, MMRM）
  (3) 解析手法の解説2（PMM, SPM）

28

6. 感度分析

感度分析

・欠測のあるデータに対する解析は仮定が多い.
→ 仮定の妥当性を検討したい
→ **「何の仮定に対する感度を見ているか」が重要**

・NRC(2010) Chapter 5 の記載. 感度分析には
(a) 完全データの分布に対する感度
(b) 外れ値・外れた症例に対する感度
(c) 欠測メカニズムに対する感度
などが考えられるが, (c)に注目.

◎ Type (i) 検証不能な仮定
Type (ii) 検証可能な仮定

◎ 感度パラメータ (sensitivity parameter)

→ (4) 欠測メカニズムに対する感度分析

Type (i) の仮定と Type (ii) の仮定

仮定には**観測データから「検証できる仮定」と「検証できない仮定」**がある．以下，用語は NRC (2010)参照．

例） ・観測されたデータの分布に正規分布を仮定
→ 観測データから妥当性が確認できる
→ Type (ii)の仮定

・欠測したデータの分布に正規分布を仮定
→ 観測データからは妥当性が確認できない
→ Type (i)の仮定

◎「どういう仮定をしているか？」によって，
「どういう感度分析が必要か？」が変わる

→ (4) 欠測メカニズムに対する感度分析

7. まとめ

本セッションの目標

欠測のあるデータの解析の

① 考え方の理解

② SASでの実行方法の理解

(a) SM, MMRM, MI, PMM, SPM
(b) 欠測メカニズムに対する感度分析

セッションの構成

(1) セッションの概要と基本事項の整理(本発表)

(2) 解析手法の解説1(選択モデル・MMRM)

(3) 解析手法の解説2
(パターン混合モデル・共有パラメータモデル)

(4) 欠測メカニズムに対する感度分析

(5) まとめと質疑応答

参考文献

1. European Medicines Agency (2010). Guideline on missing data in confirmatory clinical trials. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/09/WC500096793.pdf
2. Mallinckrodt, C. H. (2013). *Preventing and Treating Missing Data in Longitudinal Clinical Trials*. Cambridge Press.
3. National Research Council. (2010). *The Preventing and Treatment of Missing Data in Clinical Trials*. National Academies Press.
4. Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013). What Is Meant by "Missing at Random"?. *Statistical Science*, **2**, 257-268.

【企画セッション】欠測のあるデータに対する各種解析手法と欠測メカニズムに対する感度分析

(2)解析手法の解説1(SM, MMRM)

大江基貴[1)2)] 土居正明[1)3)] 縄田成毅[1)4)]
1)日本製薬工業協会 医薬品評価委員会 データサイエンス部会 タスクフォース4 欠測のあるデータに対する解析方法論・SASプログラム検討チーム
2) 株式会社大塚製薬工場 3) 東レ株式会社 4) 杏林製薬株式会社

The review of analytical approach 1 (SM, MMRM)

Motoki Oe[1)2)], Masaaki Doi[1)3)], Shigeki Nawata[1)4)]

1) The team for statistical methodologies and SAS programming of data analysis with missing data, task force 4, data science expert committee, drug evaluation committee, Japan Pharmaceutical Manufacturers Association.
2) Otsuka Pharmaceutical Factory, Inc. 3) Toray industries, Inc.
4) Kyorin Pharmaceutical CO., LTD.

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:解析手法の解説1

Selection Model(以下, SM)及びMixed effect Models for Repeated Measures(以下, MMRM)を概説し, 解析例を示す.

キーワード:LMM, MAR, Mixed effect model, MMRM, MNAR, Repeated measure, Selection model

Contents

- 尤度を用いた方法
- Selection Model（SM）
  - MARの場合
  - MNARの場合
- MMRM（Mixed effect Models for Repeated Measures）
  - モデルの概要
  - 用語の混乱について
  - 特定が必要なもの
- シミュレーション・データの解析
  - マクロの紹介
  - 解析結果

尤度を用いた方法

- 欠測を伴うデータの尤度
  - 応答変数 $\mathbf{Y}_i = (\mathbf{Y}_i^o, \mathbf{Y}_i^m)^{\mathrm{T}}$
  - 欠測識別変数 $\mathbf{R}_i$

  } 両方の尤度の寄与を考えなくてはならない
- 完全データの尤度（Full Data Likelihood）

$$L^*(\boldsymbol{\theta}, \boldsymbol{\psi}) \propto \prod_{i=1}^{N} f(\mathbf{Y}_i, \mathbf{R}_i | \boldsymbol{\theta}, \boldsymbol{\psi})$$

  - 完全データの尤度は，欠測データ $\mathbf{Y}_i^m$ も含む．
- 観測データの尤度（Observed Data Likelihood）

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{N} f(\mathbf{Y}_i^o, \mathbf{R}_i | \boldsymbol{\theta}, \boldsymbol{\psi})$$

$$= \prod_{i=1}^{N} \int f(\mathbf{Y}_i^o, \mathbf{Y}_i^m, \mathbf{R}_i | \boldsymbol{\theta}, \boldsymbol{\psi}) d\mathbf{Y}_i^m$$

  - 尤度を用いた方法では，観測データの尤度に基づいて推測を行う．

SM（Selection Model）

- SM（Selection Model）とは
  - 完全データの尤度が，以下のように分解されることを想定．

$$f(\mathbf{Y}_i, \mathbf{R}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{Y}_i|\boldsymbol{\theta}) \cdot f(\mathbf{R}_i|\mathbf{Y}_i, \boldsymbol{\psi})$$
$$= \underbrace{f(\mathbf{Y}_i^o, \mathbf{Y}_i^m|\boldsymbol{\theta})}_{\text{Type (i) の仮定}} \cdot \underbrace{f(\mathbf{R}_i|\mathbf{Y}_i^o, \mathbf{Y}_i^m, \boldsymbol{\psi})}_{\text{Type (i) の仮定}}$$

応答変数の分布のパラメータ

  - 第2項が，「観測された集団」または「欠測した集団」への個人の選択をモデル化していると解釈できるため，“Selection Model”と呼ばれる．

- SMにおける観測データの尤度

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{N} \int f(\mathbf{Y}_i^o, \mathbf{Y}_i^m|\boldsymbol{\theta}) \cdot f(\mathbf{R}_i|\mathbf{Y}_i^o, \mathbf{Y}_i^m, \boldsymbol{\psi}) d\mathbf{Y}_i^m$$

MARの場合のSM

- 観測データの尤度

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{N} \int f(\mathbf{Y}_i^o, \mathbf{Y}_i^m|\boldsymbol{\theta}) \cdot f(\mathbf{R}_i|\mathbf{Y}_i^o, \underline{\mathbf{Y}_i^m}, \boldsymbol{\psi}) d\mathbf{Y}_i^m$$
$$= \prod_{i=1}^{N} \int f(\mathbf{Y}_i^o, \mathbf{Y}_i^m|\boldsymbol{\theta}) \cdot f(\mathbf{R}_i|\mathbf{Y}_i^o, \boldsymbol{\psi}) d\mathbf{Y}_i^m$$
$$= \prod_{i=1}^{N} \{f(\mathbf{Y}_i^o|\boldsymbol{\theta}) \cdot f(\mathbf{R}_i|\mathbf{Y}_i^o, \boldsymbol{\psi})\}$$

MARの定義
$f(\mathbf{R}_i|\mathbf{Y}_i, \boldsymbol{\psi}) = f(\mathbf{R}_i|\mathbf{Y}_i^o, \boldsymbol{\psi})$

- 観測データの対数尤度

$$l(\boldsymbol{\theta}, \boldsymbol{\psi}) = \sum_{i=1}^{N} \log f(\mathbf{Y}_i^o|\boldsymbol{\theta}) + \sum_{i=1}^{N} \log f(\mathbf{R}_i|\mathbf{Y}_i^o, \boldsymbol{\psi})$$

$\boldsymbol{\theta}$ の推定に必要な部分

MARの場合のSM

- Ignorability
  - 尤度の枠組みで推測を行うことを前提として，

MARのもとでは，θ に関する推測を，
欠測過程を含めない尤度 $f(\mathbf{Y}^o|\theta)$ に基づいて行うことができる．
直接尤度（Direct Likelihood：DL）と呼ばれる

  - 欠測データを「無視」するという意味ではないことに注意．

- 他の接近法でのIgnorability
  - Bayes流接近法でも同様に成り立つ．
  - 頻度流の接近法（最小2乗法，GEEなど）では，MCARの場合でないと成り立たない（Verbeke & Molenberghs, 1997）．

MNARの場合のSM

- 観測データの対数尤度

$$l(\boldsymbol{\theta},\boldsymbol{\psi}) = \sum_{i=1}^{N} \log \underbrace{\left(\int f(\mathbf{Y}_i^o, \mathbf{Y}_i^m|\boldsymbol{\theta}) \cdot f(\mathbf{R}_i|\mathbf{Y}_i^o, \mathbf{Y}_i^m, \boldsymbol{\psi}) d\mathbf{Y}_i^m \right)}$$

θ を推定するためには，積分計算が必要

測定過程に関する尤度 $f(\mathbf{Y}_i^o, \mathbf{Y}_i^m|\boldsymbol{\theta})$

欠測過程に関する尤度 $f(\mathbf{R}_i|\mathbf{Y}_i^o, \mathbf{Y}_i^m, \boldsymbol{\psi})$

Ignorableではない！

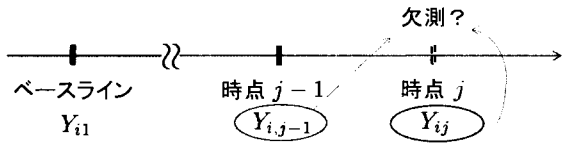
両方をモデル化しなくてはならない

MNARの場合のSM

- 欠測過程のモデル化(Type(i)の仮定)
  - ここでは, Diggle & Kenward (1994) のモデルを紹介する.
  - 単調な欠測, ベースライン測定値は欠測がないことを仮定.

$$\begin{aligned}&\text{logit}\{\Pr(R_{ij}=0|R_{i1}=1,\cdots,R_{i,j-1}=1,\mathbf{Y}_i,\boldsymbol{\psi})\}\\&\quad=\psi_0+\psi_1 Y_{i,j-1}+\psi_2\, Y_{ij}\end{aligned}$$

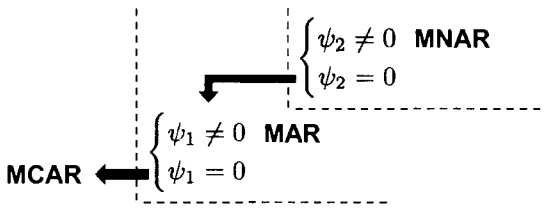
  - 時点 j における測定値が欠測であるか否かが, その1時点前の測定値とその時点の測定値に依存して決まるモデル



MNARの場合のSM

- モデルの解釈

$$\begin{aligned}&\text{logit}\{\Pr(R_{ij}=0|R_{i1}=1,\cdots,R_{i,j-1}=1,\mathbf{Y}_i,\boldsymbol{\psi})\}\\&\quad=\psi_0+\psi_1 Y_{i,j-1}+\psi_2\, Y_{ij}\end{aligned}$$



- 感度分析との関係
  - モデルを信じる立場 : ψ_0, ψ_1, ψ_2 を推定
  - Type(i)の仮定であることを強調 : ψ_2 は解析者が設定

MMRMの位置づけ

- SMの一つの形式
  - MARを仮定したもとで，DLに基づいて推測を行うことができる．
- SMの中での位置づけ

$$f(\mathbf{Y}_i, \mathbf{R}_i|\mathbf{X}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{Y}_i|\mathbf{X}_i, \boldsymbol{\theta}) f(\mathbf{R}_i|\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\psi})$$

MCAR $= f(\mathbf{R}_i|\mathbf{X}_i, \boldsymbol{\psi})$

MAR $= f(\mathbf{R}_i|\mathbf{Y}_i^o, \mathbf{X}_i, \boldsymbol{\psi})$

MNARの場合は，そのまま

欠測過程を含まない観測データの尤度(DL)

$f(\mathbf{Y}_i^o|\mathbf{X}_i, \boldsymbol{\theta})$

最大化により，θ の一致推定値を得ることができる．

LMMを想定する → MMRM

LMM(Linear Mixed Model)

- LMM(Linear Mixed Model) (Laird & Ware, 1982)

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \qquad i: \text{被験者を表す添え字}$$

  - ここに，$\mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{D})$ ，$\boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}, \Sigma_i)$
  - $\mathbf{D}$ は，変量効果間の共分散行列
  - Σ_i は，被験者 i の誤差の共分散行列

- 一般的な欠測を伴うデータ(n_i 個の繰り返し測定)の解析では...

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{1}_{n_i} b_i + \boldsymbol{\epsilon}_i$$

  - ここに，$b_i \sim \mathrm{N}(0, \nu^2)$ ，$\boldsymbol{\epsilon}_i \sim \mathrm{N}(0, \sigma^2\mathbf{I}_{n_i})$
    - 変量効果は1変量(被験者)のみ
    - 誤差は被験者間で共通かつ，測定値間で独立と想定

MMRMの位置づけ

- LMMの周辺モデル

$$\mathbf{Y}_i \sim \mathrm{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}), \quad \mathbf{V} = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^{\mathrm{T}} + \Sigma_i$$

- 周辺モデルから示唆される重要な特徴
  - 変量効果のバラつきを，周辺分散 $\mathbf{V}$ の一部と解釈することができる
    - $\mathbf{V}$ の構造としてまとめてパラメータ化すれば，変量効果を明示的にモデリングせずともよい．
    - 周辺モデルでは，多変量正規分布が想定される．

- Mallinckrodt et al.(2001)は，上記のように解釈したLMMを“MMRM”と呼んだ．

一般的なSMとMMRMの違い

- 簡単のため，balancedデータで $n_i = 3$ とする．

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{1}_3 b_i + \boldsymbol{\epsilon}_i \qquad b_i \sim \mathrm{N}(0, \nu^2), \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_3)$$

- 周辺分散 $\mathbf{V}$

$$\begin{aligned}\mathbf{V} &= \mathbf{1}_3\nu^2\mathbf{1}_3^{\mathrm{T}} + \sigma^2\mathbf{I}_3 \\ &= \nu^2\begin{pmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{pmatrix} + \sigma^2\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix} \\ &= \begin{pmatrix}\nu^2+\sigma^2 & \nu^2 & \nu^2\\ \nu^2 & \nu^2+\sigma^2 & \nu^2\\ \nu^2 & \nu^2 & \nu^2+\sigma^2\end{pmatrix}\end{aligned}$$

Compound Symmetry

MMRM

周辺分散(誤差分散)を直接にパラメータ化

$$\begin{pmatrix}\sigma_1^2 & \sigma_{12} & \sigma_{13}\\ \sigma_{12} & \sigma_2^2 & \sigma_{23}\\ \sigma_{13} & \sigma_{23} & \sigma_3^2\end{pmatrix}$$

例えば，Unstractured

用語の混乱

- “Mixed Model”という言葉を含んでいるが・・・
  - 変量効果(変量切片)を明示的に指定しない.
  - SASによる実装で, 一般にRANDOM STATEMENTは使わない.
  - ただし, 他のCluster(ex. region)は変量効果として組み込むことができる.
- どこからどこまでが“MMRM”か?
  - 額面どおりに用語を解釈すると, “MMRM”とは解析モデルの大きなクラスの一つと考えることができる.
  - Mallickrodt et al. (2001)が示したモデルは“MMRM”のひとつに過ぎず, さらに言うとこれは解析方法を示すものではない.
  - 実務的には, より詳細な仕様の特定が必要

特定が必要なもの:平均構造

- 応答と固定効果の関係
  - 欠測を伴うかどうかに関わらず, 一般の回帰モデルで明示的に特定される.
  - 一般的には, 以下のような想定が多い(と思われる)

$$\text{応答} = \text{ベースライン} + \text{治療} + \text{時点} + \text{治療}\times\text{時点}$$

  - ベースライン×時点を組み込むかどうかなど, 領域によってはさらなる議論がある(Dinh & Yang, 2011)
- 感度分析の立場
  - Restrictive Model :試験デザインに関する因子のみ(+重要な共変量)
  - Inclusive Model :脱落に関連する多くの共変量(補助変数)を含める

Inclusive Model

- Inclusive Model：補助変数を組み込んだモデル
  - 欠測がMNARであっても，適切な補助変数をモデルに組み込めば，解析上，欠測はMARに近づく．
  - Randomize後の変数は，補助変数として有用な情報をもつことが多いが，治療と交絡するおそれがあるため，一般に解析モデルに組み込むことは出来ない（ICH E9）
- 感度分析としての利用
  - 解析モデルには組み込めないが，MI（Multiple Imputation）やwGEE（weighted GEE）ならば，以下で利用できる．
    - MI　：補完モデル
    - wGEE：IPW（Inverse Probability Weighting）モデル
  - 解析モデルは，Restrictive Model
  - MARかどうかの感度分析に有用な情報を与える可能性がある．

特定が必要なもの：周辺モデルの分散共分散構造

- 様々な構造を選択することができる．
  - CS，AR(1)，Toeplitz，UNなど
- 誤特定の問題
  - MARのもとでのDLに基づく推測の妥当性には関わらないが・・・
  - 共分散構造を誤って特定すると，一般に推定量の一致性が失われる．
- 誤特定に対する処方
  - サンドウィッチ分散（ロバスト分散）の利用
    - 漸近不偏な分散推定量だが，欠測がMCARでない限りパラメータの点推定値はバイアスをもつことがある（Lu & Mehrotra, 2009）．
    - SASでは，併用可能な自由度の計算法が限られる．
  - UN：無構造（Unstractured）を指定する．
    - （多少）推定効率を犠牲にし，収束に問題を抱えることがある．

特定が必要なもの：周辺モデルの分散共分散構造

- UNを指定した場合に収束しなかったら
  - Newton-Raphson法の初期値を，Fisher's score法で与える．
    - 「初期値」である点に注意．Fisher's score法で推定を行ってはならない．
    - 欠測を伴う場合，期待情報量からは分散の一致推定量が導かれない（Verbeke & Molenberghs, 1997）．
  - 別の推定アルゴリズムを利用する．
    - 逐次単回帰法（Lu & Mehrotra, 2009）
- それでも収束しなかったら
  - 共分散構造を少しずつ特定する．
    - 特定の順序（例えば，Toeplitz⇒HCS⇒AR（1）⇒CS⇒VC）は，事前に決めておく．

収束に失敗した際のこれらの処方は．あらかじめ取り決めておくことが推奨される．

特定が必要なもの：自由度の計算法

- 自由度の計算の必要性
  - 欠測を伴い，データがUnbalancedの場合は，検定統計量（F統計量）の分母の自由度が一意に定まらない．
  - SASでは，いくつかの計算方法を選択することができる．
    - 文献で多く適用が見られるのは，SATTERTHWAITE と KENWARDROGER

- KENWARDROGER：KR法（Kenward & Roger, 1997）
  - パラメータのモデル分散（漸近分散）は，それ自体に分散共分散行列の推定量を含むため，その推定に伴うバラつきを考慮しない場合にバイアスをもつことが知られている．
  - KR法では，このバイアスを調整したモデル分散を利用して SATTERTHWAITEの近似法を利用し，分母の自由度を計算する．

MMRMで特定が必要なもの：まとめ

- 平均構造
  - 変量効果を組み込むかどうかも含む
- 推定方法：制限付き最尤法（REML）が第一選択
- （周辺モデルの）分散共分散構造
  - 収束しなかった場合の対応（構造の特定順など）
- 自由度の計算方法

これらは，統計解析計画書に事前明記することが望ましい

SASによる実装

- MMRM
  - PROC MIXEDを利用して，簡単に実装することができる．
  - プログラム例

```
PROC MIXED DATA = インプットDS名;
  CLASS      治療 時点 被験者ID;
  MODEL      応答 = ベースライン 治療 時点 治療*時点 / DDFM=KR;
  LSMEANS    治療*時点;
  REPEATED   時点 / SUBJECT=被験者ID TYPE=UN;
RUN;
```

- SM
  - MNARの場合のSMを実装するためのマクロが，DIA working groupにより公開されている．
  - ここでは，Type(i)の仮定の感度分析に使用するマクロを紹介する．

マクロの紹介（SM, Type(i)の仮定の感度分析）

- 欠測過程のモデル
  - Diggle & Kenward (1994) のモデルを利用
    - 群ごとに異なるパラメータを設定

実薬群 : $\mathrm{logit}\{\Pr(\mathrm{missing})\} = \psi_1 + \psi_3 Y_{i,j-1} + \psi_5 Y_{ij}$

プラセボ群 : $\mathrm{logit}\{\Pr(\mathrm{missing})\} = \psi_2 + \psi_4 Y_{i,j-1} + \psi_6 Y_{ij}$

  - 欠測メカニズムに応じて，モデルを選択する仕様
    - MCAR, MAR, MNARに加えて，“MNARS”を選択できる.
    - モデルはMNARと同じで，ψ_5, ψ_6 のみ固定値を割り当てることができる.

感度パラメータ

“MNARS”
DIAマクロ中の固有の名称.
“S”は，“Special case”を表す.

- 解析モデル
  - 平均構造は，MMRMと同じ指定とする.
  - 非線形最適化により，数値的に最尤推定値を得る.
    - パラメータの初期値には，MMRMの推定値を利用

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

マクロの紹介（SM, Type(i)の仮定の感度分析）

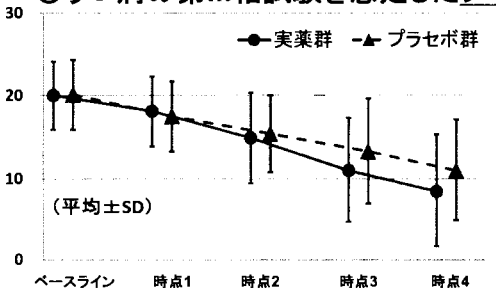
```
%macro SM_GridSearch(
        psi5grid        = 1番目のdrugの感度パラメータの指定(-1～1の範囲)
        psi6grid        = 2番目のdrugの感度パラメータの指定(-1～1の範囲)
        INPUTDS         = インプットDS名,
        COVTYPE         = 共分散構造(UN, TOEP, TOEPH, ARH, AR, CSH, CS)
        response        = 応答変数,
        MODL            = 平均構造の指定(MMRMの指定と合わせる),
        CLASVAR         = カテゴリカル変数,
        mech            = MNARS,
        const           = 数値積分の積分範囲に関係する値(3～8),
        derivative      = 数値計算に関するフラグ(0, 1),
        method          = 非線形最適化の方法
                          (NR:Newton Raphson ridge, QN:準ニュートン法),
        out1            = アウトプットDS名(パラメータの推定値),
        out2            = アウトプットDS名(差のLSMEAN),
        out3            = アウトプットDS名(LSMEAN),
        DEBUG           = 0
);
```

マクロの紹介(SM, すべてのパラメータを推定)

```
%Selection_Model2(
    INPUTDS     = インプットDS名,
    COVTYPE     = 共分散構造(UN, TOEP, TOEPH, ARH, AR, CSH, CS)
    response    = 応答変数,
    MODL        = 平均構造の指定(MMRMの指定と合わせる),
    CLASVAR     = カテゴリカル変数,
    mech        = 欠測メカニズム(MCAR, MAR, MNAR)
    const       = 数値積分の積分範囲に関係する値(3～8),
    derivative  = 数値計算に関するフラグ(0, 1),
    method      = 非線形最適化の方法
                  (NR:Newton Raphson ridge, QN:準ニュートン法),
    out1        = アウトプットDS名(パラメータの推定値),
    out2        = アウトプットDS名(差のLSMEAN),
    out3        = アウトプットDS名(LSMEAN),
    DEBUG       = 0
);
```


解析対象データ

◎うつ病の第III相試験を想定したシミュレーションデータ



●主要評価項目:HAM-D
→ スコア低下:改善
(解析には変化量使用)

●実薬群 vs プラセボ群
・1群100例(ベースライン時)

◎単調な欠測のみ

	ベースライン		時点1		時点2		時点3		時点4	
	例数	平均(SD)	例数	平均(SD)	例数	平均(SD)	例数	平均(SD)	例数	平均(SD)
実薬群	100	20.0 (4.1)	93	18.1 (4.2)	89	14.9 (5.4)	84	11.0 (6.3)	83	8.5 (6.8)
プラセボ群	100	20.1 (4.2)	90	17.5 (4.2)	87	15.4 (4.6)	85	13.3 (6.3)	80	11.0 (6.1)

26

解析結果1

psi5grid = ψ_5
psi6grid = ψ_6

- マクロ "SM_GridSearch" での解析結果
  - ここでは，Δ = psi5grid = psi6grid = 0 を指定（MARを仮定）

時点4の推定値（mean（SE））		群間差	群間差のSE	p 値
実薬群	プラセボ群			
-11.22(0.69)	-8.96(0.70)	-2.26	0.98	0.022

- MMRMでの解析結果
  - PROC MIXEDによる実装，解析モデルは“SM_GridSearch”と同じ

時点4の推定値（mean（SE））		群間差	群間差のSE	p 値
実薬群	プラセボ群			
-11.22(0.69)	-8.97(0.70)	-2.26	0.99	0.024

  - 感度パラメータ = 0 のSMの解析結果とほぼ一致

解析結果2

- マクロ“Selection_Model2”での解析結果
  - 感度パラメータも含めて，すべて推定
  - MNAR（Diggle & Kenwardの欠測過程モデル）を仮定して解析

時点4の推定値（mean（SE））		群間差	群間差のSE	p 値
実薬群	プラセボ群			
-11.37(0.70)	-9.25(0.70)	-2.12	1.01	0.037

  - 感度パラメータの推定値
    - 実薬群 : psi5grid = - 0.13
    - プラセボ群 : psi6grid = - 0.16
- 感度分析
  - MARを仮定した解析との違いは？
  - 感度パラメータを動かすと，解析結果はどう変わる？

参考文献 1

- Diggle, P. & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics.* **43**(1), 49-93.
- Dinh, P. & Yang, P. (2011). Handling baselines in repeated measures analysis with missing data at random. *Journal of Biopharmaceutical Statistics*, **21**, 326-341.
- Fitzmaurice, G., Davian, M., Verbek, G. & Molenberghs, G. (2008). *Longitudinal Data Analysis*. Chapman & Hall/CRC.
- Kenward, M. G. & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53**, 983-997.
- Laird, N. M. & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**(4), 963-974.
- Little, R. J. A. & Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley & Sons. New York.
- Lu K. & Mehrotra D. V. (2009). Specification of covariance structure in longitudinal data analysis for randomized clinical trials. *Statistics in Medicine*}, **29**, 474-488.

参考文献 2

- Mallinckrodt C. H., Clark W. S. & David S. R. (2001). Accounting for dropout bias using mixed-effects models. *Journal Biopharmaceutical Statistics*, **11**, 9-21.
- Mallinckrodt, C. H., Lane, P. W., Schnell, D., Peng, Y. & Maucuso, J. P. (2008). Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Durg Information Journal*, **42**, 303-319.
- Mallinckrodt, C. H. (2013). *Preventing and Treating Missing Data in Longitudinal Clinical Trials*. Cambridge University press.
- National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: The National Academies Press.
- Verbeke, G. & Molenberghs, G. (1997). *Linear mixed models in practice: a SAS oriented approach*, New York : Springer-Verlag.

ご清聴ありがとうございました

【企画セッション】欠測のあるデータに対する各種解析手法と欠測メカニズムに対する感度分析

(3)解析手法の解説2

高橋文博[1)2)] 藤原正和[1)3)] 大浦智紀[1)4)] 横山雄一[1)5)]
1) 日本製薬工業協会 医薬品評価委員会 データサイエンス部会
タスクフォース4 欠測のあるデータに対する解析方法論・SASプログラム検討チーム
2) 田辺三菱株式会社 3) 塩野義製薬株式会社
4) 日本イーライリリー株式会社 5) 持田製薬株式会社

The review of analytical approach 2 (MI, PMM, SPM)

Fumihiro Takahashi[1)2)], Masakazu Fujiwara[1)3)],
Tomonori Oura[1)4)], Yuichi Yokoyama[1)5)],
1) The team for statistical methodologies and SAS programming of data analysis with missing data, task force 4, data science expert committee, drug evaluation committee, Japan Pharmaceutical Manufacturers Association.
2) Mitsubishi Tanabe Pharma Co., Ltd., 3) Shionogi & Co., Ltd. ,
4) Eli Lilly Japan K. K., 5) Mochida Pharmaceutical Co., Ltd.,

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

多重補完法(Multiple Imputation), 制約条件を用いたPattern Mixture model, 及びShared Parameter Modelの理論的解説, 並びにシミュレーションデータに対するSASの実行結果を提示する.

キーワード:

PMM, Identifying restrictions, CCMV, ACMV, NCMV, NFMV, Multiple Imputation, Shared Parameter Model, MNAR, NLMIXED procedure

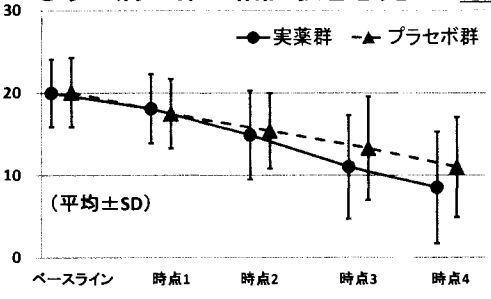
発表の流れ

1. Multiple Imputation
2. Pattern Mixture Model ➡ ※欠測メカニズムに対する感度分析セッションの前準備
   1. Pattern Mixture Modelとは
   2. 制約条件を仮定したPattern Mixture Model
      1. CCMV, NCMV, NFMV
      2. 実行方法, Multiple Imputation, マクロ
3. Shared Parameter Model
   1. Shared Parameter Modelとは
   2. 位置づけ
   3. マクロ
   4. シミュレーションデータを用いた解析例

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

解析対象データ

◎うつ病の第III相試験を想定したシミュレーションデータ



- 主要評価項目：HAM-D
  → スコア低下：改善
  （解析には変化量使用）
- 実薬群 vs プラセボ群
  ・1群100例（ベースライン時）

◎単調な欠測のみ

	ベースライン		時点1		時点2		時点3		時点4	
	例数※	平均(SD)	例数	平均(SD)	例数※	平均(SD)	例数	平均(SD)	例数	平均(SD)
実薬群	100	20.0 (4.1)	93	18.1 (4.2)	89	14.9 (5.4)	84	11.0 (6.3)	83	8.5 (6.8)
プラセボ群	※100	20.1 (4.2)	90	17.5 (4.2)	87	15.4 (4.6)	85	13.3 (6.3)	80	11.0 (6.1)

Multiple Imputation
(MI)

MARに基づく
MIの解析
(主解析)

欠測値の
データに対する
MI

➢Multiple Imputation（MI）

- Rubin（1978,1987）によって提案された方法
- 欠測メカニズムがMARであれば，MARに基づくMIの解析(主解析)は妥当
- 複数回の補完を行うことで，欠測値の補完に対して不確実性を考慮
- MIにおける解析モデルと補完モデルが尤度ベースのモデルと同じならば，MIの結果は尤度ベースの結果と類似．（Mallinckrodt, 2013）

補完モデル：欠測値を補完するための統計モデル
解析モデル：多重補完された完全データを用いて解析するための統計モデル

- MIが有用な状況（Dmitrienko et al, 2005）：
  - 共変量の欠測を補完
  - PMM（MNAR）の枠組みでも適用可能
  - 非単調な欠測を補完

➢Multiple Imputation（MI）のSASの実行種類

- 単調又は非単調 欠測パターン仮定
- 表1 Proc MIにおける補完方法 (Yang Yuan, 2011)

Pattern of missingness	Type of imputed variable	Available methods
Monotone	Continuous	Monotone regression
		Monotone predicted mean matching
		Monotone propensity score
Monotone	Classification (ordinal)	Monotone logistic regression
Monotone	Classification (nominal)	Monotone discriminant function
Arbitrary	Continuous	MCMC full-data imputation
		MCMC monotone-data imputation

Table 1: Imputation methods in PROC MI.

➢MIにおける単調回帰の概略

- 欠測値Y_jを補完するため，観測されたデータ$Y_1, ..., Y_j$を用いて補完モデルを構築
  $Y_j = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$
- 回帰パラメータの推定値$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$
- 共分散行列$\hat{\sigma}_j^2 \mathbf{V}_j$, $\mathbf{V}_j$は$(\mathbf{XX})^{-1}$より導出
- パラメータの事後予測分布から下記をサンプル
  $\beta_* = (\beta_{*0}, \beta_{*1}, \beta_{*2}, ..., \beta_{*k})$, σ_{*j}^2
  分散: $\sigma_{*j}^2 = \hat{\sigma}_{*j}^2 (n_j - k - 1)/g, g \sim \chi_{n_j - k - 1}^2$
  n_jはY_jが観測されているデータ数
  回帰係数: $\beta_* = \hat{\beta} + \sigma_{*j} \mathbf{V}_{hj}^T \mathbf{Z}$, $\mathbf{V} = \mathbf{V}_{hj} \mathbf{V}_{hj}^T$
  Z: k+1の長さの独立正規変数
- 欠測値を下記の式で補完
  $\beta_{*0} + \beta_{*1} X_1 + \beta_{*2} X_2, ..., \beta_{*k} X_k + z_i \sigma_{*j}$

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

➢単調回帰のMIによるSAS プログラム（例）

①MIの実行

```
proc mi data=XXXX
  seed=シード番号
  nimpute=補完回数（補完されたデータセットの作成個数）
  out=多重補完後の出力データセット名;
  by 薬剤群変数;
  monotone method=reg(単調回帰の場合);
  var v0 val1 val2 val3 val4（解析対象変数）;
run;
```

3回～5回が推奨
(Molenberghs et al 2007)

②多重補完後，多重補完によって得られた複数のデータセットのそれぞれに対して解析モデルを適用.

例)Mixed プロシジャーによるANCOVAモデル
最終時点の値=治療効果+ベースライン値

10

➢単調回帰のMIIによるSAS プログラム(例)

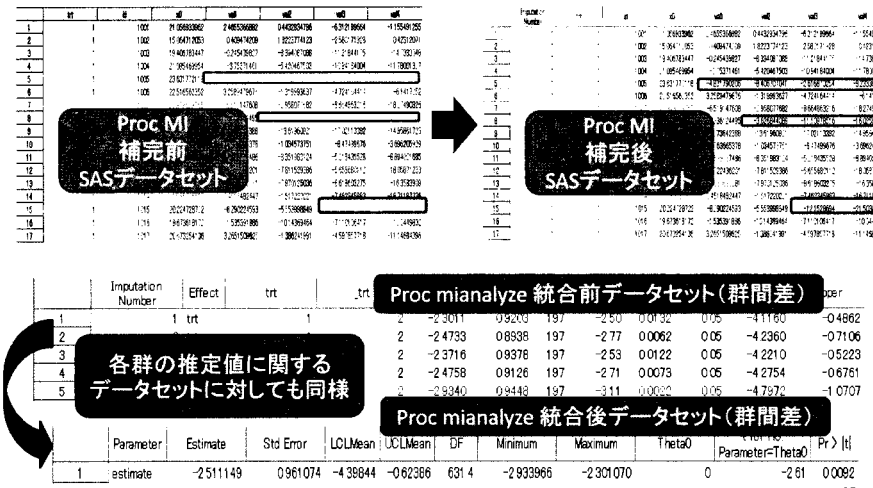
③解析モデルで得られた複数の解析結果より MIANALYZEを用いて解析結果の統合

```
proc mianalyze data=XXX alpha=0.05;
  modeleffects estimate(対象推定値の変数);
  stderr stderr(標準誤差の変数);
  ods output ParameterEstimates=統合結果;
run;
```


医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

➢単調回帰のMIIによるSAS 解析結果(例)

①及び③で得られるSASデータセット



	Imputation Number	Effect	trt	_trt	DF	tValue	Probt	Estimate	Std Error	LCLMean	UCLMean
1	1	trt	1		197	-2.50	0.0132	0.05	-4.1160	-0.4862	
2	2				197	-2.77	0.0062	0.05	-4.2360	-0.7106	
3	2				197	-2.53	0.0122	0.05	-4.2210	-0.5223	
4	2				197	-2.71	0.0073	0.05	-4.2754	-0.6761	
5	2				197	-3.11	0.0022	0.05	-4.7978	-1.0707	

	Parameter	Estimate	Std Error	LCLMean	UCLMean	DF	Minimum	Maximum	Theta0	Parameter=Theta0	Pr > |t|
1	estimate	-2.511149	0.961074	-4.39844	-0.62386	631.4	-2.933966	-2.301070	0	-2.61	0.0092

➢MI(単調回帰)+ANCOVA:解析対象データの解析結果

解析モデル 欠測メカニズム /欠測確率	時点4における各群の点推定値(SE)		群間差	群間差のSE	P値
	実薬群	プラセボ群			
MI MNAR/Low	-11.26(0.68)	-8.75(0.67)	-2.51	0.96	0.009

- 多重補完後の解析モデル: ANCOVA

```
proc mixed data=/*補完後データセット*/;
   by _Imputation_;
   class trt;
   model val4=x0 trt;
   lsmeans trt/pdiff=control("2") cl alpha=0.05;
   ods output diffs=/*群間差のデータセット*/;
   ods output lsmeans=/*各群の推定値のデータセット*/;
run;
```


Pattern Mixture Model
(PMM)

➢PMMとは

*Little (1993,1994,1995),Ratitch et al. (2013)

$$Pr\left(\mathrm{Y^O}, \mathrm{Y^M}, \mathrm{R} \setminus \mathrm{X}\right)$$
$$= Pr\left(\mathrm{R} \setminus \mathrm{X}\right) Pr\left(\mathrm{Y^O}, \mathrm{Y^M} \setminus \mathrm{R}, \mathrm{X}\right)$$
$$= Pr\left(\mathrm{R} \setminus \mathrm{X}\right) Pr\left(\mathrm{Y^O} \setminus \mathrm{R}, \mathrm{X}\right) Pr\left(\mathrm{Y^M} \setminus \mathrm{Y^O}, \mathrm{R}, \mathrm{X}\right)$$

Selection model

$$Pr\left(\mathrm{Y^O}, \mathrm{Y^M}, \mathrm{R} \setminus \mathrm{X}\right)$$
$$= Pr\left(\mathrm{R} \setminus \mathrm{Y^O}, \mathrm{Y^M}, \mathrm{X}\right) Pr\left(\mathrm{Y^O}, \mathrm{Y^M} \setminus \mathrm{X}\right)$$

観測データ，欠測データ，欠測パターン変数の同時分布の分解方法が異なる

推定不可能な欠測データの確率分布に何らかの制約条件が必要

$\mathrm{Y^O}$: 観測されたアウトカム

$\mathrm{Y^M}$: 観測されなかった（欠測）アウトカム

R : 欠測識別変数

X : 観測されている共変量

NRC [2010]
Type 1: 欠測データの分布に対する検証不可能な仮定
Type 2: 観測データの分布に対する検証可能な仮定

観測データから推定可能 Type (i)

観測データから推定不可能 Type (ii)

➢欠測パターンの例 欠測メカニズムがMARやMNARの状況を想定

①脱落時点に基づく脱落パターン

例) 観測○ 未観測×

時点1 ○	時点2 ○	時点3 ○	時点4 ○	パターン1
時点1 ○	時点2 ○	時点3 ○	時点4 ×	パターン2
時点1 ○	時点2 ○	時点3 ×	時点4 ×	パターン3
時点1 ○	時点2 ×	時点3 ×	時点4 ×	パターン4

②中止理由に基づく脱落パターン

中止理由1：有害事象により中止

中止理由2：効果不足

中止理由3：被験者の同意撤回 など

他にもパターンの定義は可能

(ア)パターン1のデータのみで解析した場合，バイアスが入る
(イ)補完してバイアスを減らしたい
(ウ)制約条件を仮定し，×のデータを補完
=>×の分布が不明
=>制約条件（欠測メカニズムと対応）を仮定し，周辺の○のデータを参考に，×のデータを補完

➢ 制約条件と欠測メカニズムの関係図

Missing data in clinical trials page 37, Molenberghs et al. (2007)より

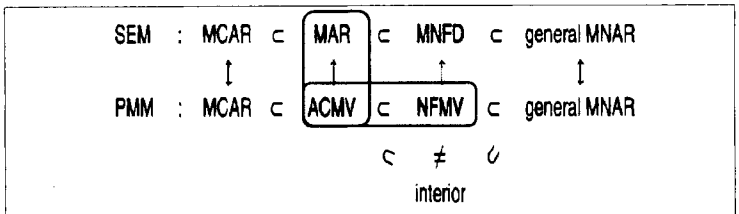


Figure 3.1 Relationship between nested families within the selection model (SEM) and pattern-mixture model (PMM) families. (MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random; MNFD, missing non-future dependence; ACMV, available case missing values; NFMV, non-future missing values; interior, restrictions based on a combination of the information available for other patterns. The ⊂ symbol here indicates 'is a special case of'. The ↕ symbol indicates correspondence between a class of SEM models and a class of PMM models.)

17

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

➢制約条件(CCMV,NCMV,ACMV)を仮定したPMM

Thijs et al. (2002)は上記の制約条件を統一的に示す形でPMMを提案

t時点まで観測

$$f(y_1,...,y_T,\boxed{r=t}) = f_t(y_1,..,y_t)\boxed{f_t(y_{t+1},y_{t+2},...,y_T|y_1,..,y_t)}$$

未観測データの分布

脱落パターン $r = 1,\cdots,T$

➢制約条件(CCMV,NCMV,ACMV)

$$f_t(y_s|y_1,...,y_{s-1}) = \sum_{j=s}^{T}\boxed{\omega_{sj}}f_j(y_s|y_1,...,y_{s-1}),\ s = t+1,..,T$$

この重みω_{sj} (s:補完対象時点, j:どのパターンの分布)の設定により, 制約条件CCMV,NCMV,ACMVを表現

以上, 3つの制約条件を完全データの分布関数に代入

$$f_t(y_1,....,y_T) = f_t(y_1,..,y_t)\prod_{s=0}^{T-t-1}\left(\sum_{j=T-s}^{T}\omega_{T-s,j}f_j(y_{T-s}|y_1,....,y_{T-s-1})\right)$$

18

➢制約条件1(CCMV:Complete Case Missing Value)

$\omega_{T,T} = \omega_{T-1,T} = \omega_{T-2,T} = \ldots = \omega_{t+1,T} = 1$
かつその他 $\omega_{sj} = 0, j \neq T$ のとき
$f_t(y_s|y_1, \ldots, y_{s-1}) = f_T(y_s|y_1, \ldots, y_{s-1}), s = t+1, \ldots, T$

最後まで観測されたCompleter（例，パターン1）の情報から欠測の情報を補完するもの．完全パターンのデータを利用し，多くの被験者がCompleterのパターンの場合を満たしている場合に有用．また，Non-monotoneのときにも利用が容易．

例） 観測○ 未観測×

パターン1 時点1 ○ 時点2 ○ 時点3 ○ 時点4 ○
パターン2 時点1 ○ 時点2 ○ 時点3 ○ 時点4 ×
パターン3 時点1 ○ 時点2 ○ 時点3 × 時点4 ×

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

➢制約条件2(NCMV:Neighboring Case Missing Values)

$\omega_{T,T} = \omega_{T-1,T-1} = \omega_{T-2,T-2} = \ldots = \omega_{t+1,t+1} = 1$
かつその他 $\omega_{sj} = 0, j \neq s$ のとき
$f_t(y_s|y_1, \ldots, y_{s-1}) = f_s(y_s|y_1, \ldots, y_{s-1}), s = t+1, \ldots, T$

欠測時点で測定された一番近いパターンの情報から欠測の情報を補完するもの．たとえば，下記の例でパターン3の時点3を補完したい場合，一番欠測パターンが近いパターン2のデータの分布から補完

例） 観測○ 未観測×

パターン1 時点1 ○ 時点2 ○ 時点3 ○ 時点4 ○
パターン2 時点1 ○ 時点2 ○ 時点3 ○ 時点4 ×
パターン3 時点1 ○ 時点2 ○ 時点3 × 時点4 ×

20

➢制約条件(NFMV)を仮定したPMM

Kenward et al. (2003)はMNARを仮定したPMMを提案
ACMVの結果とNFMV+CCMV又はNFMV+NCMVの結果比較することは,
MARの欠測メカニズムからの乖離を評価

$$
\begin{aligned}
f(y_1, \ldots y_T, r=t) &= f(y_1, \ldots, y_T | r=t) f(r=t) \\
&= f_t(y_1, \ldots, y_T) f(r=t) \\
&= f_t(y_1, \ldots, y_t) f_t(y_{t+1} | y_1, \ldots, y_t) f_t(y_{t+2}, \ldots, y_T | y_1, \ldots, y_{t+1}) \\
&\quad \times f(r=t)
\end{aligned}
$$

観測

$f_t(y_1, \ldots, y_t)$ $f_t(y_{t+1} | y_1, \ldots, y_t)$ $\prod_{s=t+2}^{T} f_t(y_s | y_1, \ldots, y_{s-1})$

過去(t時点まで) 現在(t+1時点) 未来(t+2時点以降)

観測 未観測 未観測

制約条件が必要

21

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

➢制約条件(NFMV: Non-Future Missing Value)

$s \geq t+2$ に対して
$f(y_s | y_1, \ldots, y_{s-1}, r=t) = f(y_s | y_1, \ldots, y_{s-1}, r \geq s-1)$

- 脱落が今現在観測されていないデータに依存し,
  未来に観測されていないデータには依存しない
- 欠測メカニズム(Selection ModelのMissing Non-Future Dependent:MNFD)に対応

現在(t+1)の未観測のデータ

$f(r=t | y_1, \ldots, y_T) = f(r=t | y_1, \ldots, y_{t+1})$

- 完全にデータの分布が特定できない, すなわち下記の
  条件付き分布のみが特定されていない
  =>制約条件(CCMVやNCMV)の追加が必要

$f(y_s \setminus y_1, \ldots, y_{s-1}, r=s-1)$ NFMV+CCMV,NFMV+NCMV => MNARを仮定

22

➢PMMマクロ（参考：Missingdata.org.uk）

```
%macro %patternmixture(analyset=XXX, /*データセット名*/
lconstraint=XXX, /*補完モデルに使用, CCMV or NCMV*/
ltype=XXX, /*時点制約としてNFMV or ALL*/
seedgen=XXX, numberimputations=X, /*シード数と補完数*/
YVAR=%STR(XXX), /*目的変数*/
MODEL=%STR(XXXX), /*補完モデル*/
modvars=%STR(xxx),
classvars=%STR(XXX), id=, /*補完モデルの質的変数*/
MODEL2=%STR(XX), /*解析モデル, 補完モデルと同一でなくてよい*/
classvars2=%STR(XX), /*解析モデルの質的変数*/
othervars=%STR(XXX), /*Lsmean算出に必要な連続変数*/
TITLE=%STR(),
FOOTNOTE=%STR( ));
```


➢PMMマクロの例

- NFMV+CCMV

補完モデル：
治療効果，時点効果，ベースライン値の飽和モデル
解析モデル：
ベースライン値，脱落パターン，治療，時点，脱落パターンと治療の交互作用，治療と時点の交互作用，脱落パターンと治療と時点の交互作用
制約条件：NFMV+CCMV
データセット：MNARかつ欠測確率がLow

```
* complete case;
%patternmixture(
  analyset=D7,
  lconstraint=CCMV,
  ltype=NFMV,
  seedgen=8857954,
  numberimputations=5,
  YVAR=%STR(VAL),
  MODEL=%STR(base treatment visit base*treatment base*visit treatment*visit base*treatment*visit),
  modvars=%STR(subjid treatment base),
  classvars=%STR(treatment visit), id=,
  MODEL2=%STR(base dgroup treatment visit dgroup*treatment treatment*visit dgroup*treatment*visit),
  classvars2=%STR(dgroup treatment visit),
  othervars=%STR(base),
  TITLE=%STR(CCMV Pattern Mixture Model Analysis),
  FOOTNOTE=%STR( )
);
```

➢PMMマクロ（データ：MNAR，欠測確率Low）のSAS output

- NFMV+CCMV

CCMV Pattern Mixture Model Analysis
Using CCMV identifiability constraint
By visit LSMEANS using Model with Effects for Dropout Pattern

Visit (Week)	Treatment	N	LSMEAN Change	SE	LSmean Difference	SE	Lower CL	Upper CL	P-value
Visit 1	Drug	93	-1.97	0.44					
	Placebo	90	-2.46	0.44	0.49	0.63	-0.74	1.72	4329
Visit 2	Drug	89	-5.01	0.52					
	Placebo	87	-4.58	0.53	-0.42	0.73	-1.86	1.01	5621
Visit 3	Drug	84	-8.72	0.68					
	Placebo	85	-6.59	0.67	-2.13	0.97	-4.04	-0.22	0288
Visit 4	Drug	83	-11.28	0.74					
	Placebo	80	-8.85	0.72	-2.43	1.02	-4.43	-0.43	0171
Overall	Drug	93	-6.75	0.49					
	Placebo	90	-5.62	0.49	-1.12	0.69	-2.48	0.23	1046

NFMV observations used subject to identifiable constraint
Model base dgroup treatment visit dgroup*treatment treatment*visit dgroup*treatment*visit

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

Shared Parameter Model (SPM)

Shared Parameter Model

$$f(Y_i, R_i, b_i) = f(Y_i|R_i, b_i)f(R_i|b_i)f(b_i)$$

i 被験者　　$\mathbf{b}_i = (b_{i1}, \ldots, b_{iq})$　　$\mathbf{b}_i \sim N(\mathbf{0}, \Sigma)$

- 測定過程に対するモデル，及び脱落過程に対するモデルの両方に影響する潜在変数 (変量効果) を考える.
- 変量効果の条件付きで，測定過程と脱落過程を表す密度関数は分離できると仮定する.

定義

$$f(Y_i, R_i, b_i) = f(Y_i^o|b_i)f(Y_i^m|b_i)f(R_i|b_i)f(b_i)$$

Shared Parameter Model

$$f(Y_i, R_i, b_i) = f(Y_i^o|b_i)f(Y_i^m|b_i)f(R_i|b_i)f(b_i)$$

- ある時点における脱落が，アウトカムではなく，アウトカムにも関連する個々人の潜在的な特性の影響を受けると考える. (Little, 1995)
- 共通の変量効果が測定過程モデルと脱落過程モデルに含まれると考える場合，欠測メカニズムがMNARの場合に対応する.
- SM(Selection Model)との違い

$$f(Y_i, R_i) = f(Y_i^o, Y_i^m)f(R_i|Y_i^o, Y_i^m)$$

尤度の計算

$$
\begin{aligned}
f(Y_i^o, R_i) &= \int_{y^m}\int_b f(Y_i^o, Y_i^m, r_i, b_i)db_i dY_i^m \\
&= \int_{y^m}\int_b f(Y_i^o, Y_i^m|b_i)f(r_i|b_i)f(b_i)db_i dY_i^m \\
&= \int_{y^m}\int_b f(Y_i^o|b_i)f(Y_i^m|b_i)f(r_i|b_i)f(b_i)db_i dY_i^m \\
&= \int_b f(Y_i^o|b_i)f(r_i|b_i)f(b_i)\left(\int_{y^m} f(Y_i^m|b_i)dY_i^m\right)db_i \\
&= \int_b f(Y_i^o|b_i)f(r_i|b_i)f(b_i)db_i
\end{aligned}
$$

SAS のNLMIXED プロシジャを用いることで実装可能である.

Shared Parameter Modelの位置づけ

- 感度分析としてのSPM
  - 欠測がMNARの場合において, Selection Model, 及び Pattern Mixture Modelと並ぶ感度分析の方法である.
    - アウトカムの測定誤差が大きく, 脱落するかどうかにアウトカムの値が依存すると考えるより, 症状や病態の進行度合いのような個々人の潜在的な効果が関係していると思われる状況において, 有用なアプローチと考えられる. (Little, 1995)

SPM $f(Y_i, R_i, b_i) = f(Y_i^o|b_i)f(Y_i^m|b_i)f(R_i|b_i)f(b_i)$

SM $f(Y_i, R_i) = f(Y_i^o, Y_i^m)f(R_i|Y_i^o, Y_i^m)$

PMM $f(Y_i, R_i) = f(Y_i^o, Y_i^m|R_i)f(R_i)$

マクロ:%shared_parameter

- DIA working group公開マクロを参考にして作成

```
%shared_parameter(
INPUTDS = temp_sim1,                          インプットデータ
SUBJVAR = id,                                 被験者
TRTVAR = Drug,                                薬剤
TIME = time,                                  時点
MODL =
   %STR(val = x0 Drug time Drug*time),        測定過程に対するモデル
                                              に含まれる固定効果

LINK=%STR(CLOGLOG)                            脱落仮定に対するモデル

RANDOM_SLOPE =%STR(LINEAR)                    変量効果の選択
);
```


マクロの仕様

- マクロでは4パターンのモデルが表現可能

モデル1	モデル2
測定過程のモデルに含めた変量効果を，脱落過程モデルは含まない	測定過程のモデルに含めた変量効果を，脱落過程モデルも含む

各モデルに対して‥

RANDOM_SLOPE =%STR(NONE)
　変量効果は切片しか含まない

RANDOM_SLOPE =%STR(LINEAR)
　変量効果は，切片，及び時点に対する効果も含む

マクロ:%shared_parameter

- 測定過程に対するモデル

RANDOM_SLOPE =%STR(NONE)
変量効果は切片しか含まない

$$Y_{ij} = \beta_0 + \beta_1 Time_{ij} + \beta_2 Group_i + \beta_3 Time_{ij} \times Group_i + b_{i0} + e_{ij}$$

RANDOM_SLOPE =%STR(LINEAR)
変量効果は，切片，及び時点に対する効果も含む

$$Y_{ij} = \beta_0 + \beta_1 Time_{ij} + \beta_2 Group_i + \beta_3 Time_{ij} \times Group_i + b_{i0} + b_{i1} Time_{ij} + e_{ij}$$

マクロ:%shared_parameter

- 脱落過程に対するモデル
  - Complementary log-log linkモデル

RANDOM_SLOPE =%STR(NONE)
変量効果は切片しか含まない

$$Pr(D_i = j | D_i \geq j) = 1 - exp(-exp(\alpha_{0j} + \gamma_1 b_{i0}))$$

RANDOM_SLOPE =%STR(LINEAR)
変量効果は，切片，及び時点に対する効果も含む

$$Pr(D_i = j | D_i \geq j) = 1 - exp(-exp(\alpha_{0j} + \gamma_1 b_{i0} + \gamma_2 b_{i1}))$$

- ロジットモデルも使用可能

シミュレーションデータを用いた解析例 (1)

Parameter	No Linkage ML est	std error	p-value	Intercept linkage (do not vary by treatment) ML est	std error	p-value	Intercept linkage (vary by treatment) ML est	std error	p-value
b0	14.205	1.693	<.001	14.185	1.685	<.001	14.150	1.683	<.001
bDrug	1.532	0.887	.086	1.519	0.888	.089	1.505	0.889	.092
bDrugtime	-1.050	0.251	<.001	-1.043	0.251	<.001	-1.011	0.252	<.001
btime	-2.150	0.178	<.001	-2.133	0.179	<.001	-2.146	0.179	<.001
bx0	-0.726	0.078	<.001	-0.725	0.078	<.001	-0.723	0.078	<.001
aDrug	-0.015	0.447	.972	-0.010	0.452	.982	-0.308	0.569	.589
aInt				0.112	0.073	.129	0.031	0.099	.754
aDint							0.179	0.158	.257
endpoint trt diff	-2.667	0.761	<.001	-2.652	0.762	<.001	-2.539	0.770	.001
endpoint trt diff (Mixed Model)	-2.668	0.761	<.001						

Mixed Model　モデル1　モデル2

RANDOM_SLOPE =%STR(NONE)

変量効果は切片しか含まない

35

シミュレーションデータを用いた解析例 (2)

Parameter	No Linkage ML est	std error	p-value	Intercept and Linear slope linkage (do not vary by treatment) ML est	std error	p-value	Intercept and Linear slope linkage (vary by treatment) ML est	std error	p-value
b0	13.764	1.489	<.001	12.147	1.488	<.001	12.147	1.481	<.001
bDrug	1.491	0.695	.033	0.943	0.688	.172	0.943	0.677	.166
bDrugtime	-1.022	0.281	<.001	-0.875	0.291	.003	-0.875	0.173	<.001
btime	-2.135	0.200	<.001	-2.167	0.207	<.001	-2.167	0.114	<.001
bx0	-0.705	0.070	<.001	-0.616	0.070	<.001	-0.616	0.070	<.001
aDrug	-0.015	0.447	.973	-2.080	3.004	.490	-2.080		
aInt				-14.117	7.039	.046	-14.117	1.334	<.001
aSlp				22.368	10.085	.028	22.368		
aDint							-14.117		
aDslp							22.368		
endpoint trt diff	-2.596	0.939	.006	-2.559	0.976	.009	-2.559		
endpoint trt diff (Mixed Model)	-2.596	0.939	.006						

Mixed Model　モデル1　モデル2

RANDOM_SLOPE =%STR(LINEAR)

変量効果は，切片，及び時点に対する効果も含む

36

参考文献

- Dmitrienko, A., Molenberghs, G., Chuang-Stein, C., Offen, W., (2005) Analysis of Clinical Trials Using SAS: *A Practical Guide* .Carry, NC: SAS Institute Inc.
- Kenward, M.G., Molenberghs, G. & Thijs, H. (2003) Pattern mixture models with proper time dependence. *Biometrika*. **90**(1), 53-71
- Mallinckrodt, C. H. (2013). *Preventing and Treating Missing Data in Longitudinal Clinical Trials.* Cambridge University press.
- Molenberghs, G., Kenward, M.G. (2007) Missing Data in Clinical Studies. *Chichester. UK, John Wiley*.
- National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: The National Academies Press.
- Ratitch, B., O'Kelly, M., Tosiello, R. (2013) Missing data in clinical trials from clinical assumptions to statistical analysis using pattern mixture models. *Phramaceutical Statistics*. 12, 337-347

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

参考文献

- Rubin DB (1978). Multiple Imputation in sample surveys-Aphenomenological Bayesian approach to nonresponse. Imputation and Editing of Faulty or Missing Survey Data. Washington, DC:U.S. Department of Commerce
- Rubin DB (1987). Multiple Imputation for Nonresponse in Surveys. *John Wiley & Sons*.
- Yang Yuan (2011). Multiple Imputation Using SAS Software. *Journal of Statistical Software*. **45**(6), 1-25 .
- Thijs, H., Molenberghs, G. (2002) Strategies to fit pattern mixture models. Biostatistics. **3**(2), 245-265
- Follmann, D. and Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. Biometrics 51, 151–168.
- Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. Journal of the American Statistical Association **88**, 125–134.

参考文献

- Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika* **81**, 471–483.
- Little, R. J. A. (1995). Modelling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association,* **90**, 1112–1121.
- Wu M., & Follmann, D.A. "Use of Summary Measures to Adjust for Informative Missingness in Repeated Measures with Random Effects," *Biometrics*, (1999); 55 75-84

【企画セッション】欠測のあるデータに対する各種解析手法と欠測メカニズムに対する感度分析

(4)欠測メカニズムに対する感度分析

駒嵜弘[1)2)] 高橋文博[1)3)] 横溝孝明[1)4)]
1)日本製薬工業協会 医薬品評価委員会 データサイエンス部会 タスクフォース4 欠測のあるデータに対する解析方法論・SASプログラム検討チーム
2) マルホ株式会社 3) 田辺三菱製薬株式会社 4) 大正製薬株式会社

Sensitivity analysis for the missing mechanism.

Hiroshi Komazaki[1)2)], Fumihiro Takahashi[1)3)], Takaaki Yokomizo[1)4)]
1) The team for statistical methodologies and SAS programming of data analysis with missing data, task force 4, data science expert committee, drug evaluation committee, Japan Pharmaceutical Manufacturers Association.
2) Maruho Co., Ltd. 3)Mitsubishi Tanabe Pharma Corporation 4) Taisho Pharmaceutical Co., Ltd.

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

Selection Model 及びPattern Mixture modelを用いて，欠測メカニズムに対する感度分析の方法を説明する．また,SASによるシミュレーション結果を提示する．

キーワード:欠測メカニズム，感度分析，MAR，MNAR，Type(i)・Type(ii)の仮定，感度パラメータ，Selection Model，Pattern Mixture model,

Contents

- NRC(2010)で提案された感度分析の種類
- 欠測メカニズムに対する感度分析
  - ✓ 感度分析を行う理由
    - ➢ 主要な解析はMAR？ or MNAR？
    - ➢ Type(i), Type(ii)の仮定と感度パラメータ
  - ✓ SMを用いた感度分析
    - ➢ モデルの説明，感度パラメータによるMAR,MNARの仮定
    - ➢ シミュレーションデータを用いた解析例
  - ✓ PMMを用いた感度分析
    - ➢ モデルの説明，感度パラメータによるMAR,MNARの仮定
    - ➢ シミュレーションデータを用いた解析例
  - ✓ SMとPMMの感度分析の違い

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

Contents

- 本感度分析の結果の考察方法～NRC(2010)の内容より～

記号の整理 (Selection Model)

対象となるデータ：経時データ（連続値）

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_i^o \\ \mathbf{Y}_i^m \end{pmatrix} \quad (i = 1, \cdots, N)$$

n_i ：被験者 i の（計画された）測定時点

N ：被験者数

$\mathbf{Y}_i^o$ ：観測データ

$\mathbf{Y}_i^m$ ：欠測データ

欠測識別変数

$$R_{ij} = \begin{cases} 1 & \text{被験者 } i \text{ の } j \text{ 時点でのデータが観測された} \\ 0 & \text{被験者 } i \text{ の } j \text{ 時点でのデータが欠測} \end{cases} \qquad \mathbf{R}_i = \begin{pmatrix} R_{i1} \\ \vdots \\ R_{in_i} \end{pmatrix}$$

パラメータ

$\boldsymbol{\theta}$ ：応答変数モデルのパラメータ　　$\boldsymbol{\psi}$ ：欠測モデルのパラメータ

記号の整理 (Pattern Mixture Model)

対象となるデータ：ある測定時点のデータ（連続値）

Y_i^m ：被験者iの未観測データ

Y_i^O ：被験者iの観測データ

μ_0 ：未観測データの平均

μ_1 ：観測データの平均

欠測識別変数

$$R = \begin{cases} 1 & \text{データが観測} \\ 0 & \text{データが欠測} \end{cases}$$

$\boldsymbol{\pi}$ ：観測割合

パラメータ

$\beta = (\beta_0, \beta_1)^{\mathrm{T}}$ ：応答変数モデルのパラメータ

NRC(2010)で提案された感度分析の種類

- 完全データの分布の仮定
- 外れ値の影響
- 欠測メカニズムの仮定（MAR or MNAR） ← 本セッションで取り上げる感度分析 NRC (2010)で『最も重要』と指摘

主要な解析はMAR or MNAR？

- MARを仮定した解析
  - ✓ 途中脱落が多いと想定される臨床試験では，周到に計画された上で，欠測がMARとなるよう十分な情報を収集できる試験デザイン考えるべき
  - ✓ しかしながらMARの仮定はデータからは確認できない
  - ✓ 実際のデータは欠測メカニズムがMARやMNARの混合である可能性もある
- 欠測がMNARの仮定の下で解析する際の問題点
  - ✓ 応答変数モデルまたは欠測モデルの仮定を恣意的に置く必要がある
  - ✓ 妥当性の検証できない仮定をしている
  - ✓ 仮定が誤っていた場合の結果の頑健性は脆い

＊Mallinckrodt (2013), NRC(2010)より抜粋

□欠測メカニズムに対する感度分析を含む解析手順の提案

1. 主解析はMARを仮定した解析
2. 感度分析として，MNARの仮定の下で解析
3. MARを仮定した解析の結果の頑健性を確認し，試験結果の妥当性を述べる．

□ これら感度分析はSM及びPMMより実行可能．

		欠測メカニズム		
		MAR		MNAR
応答変数モデル	SM	MMRM ≒	SM (MAR) ⟷	SM (MNAR)
	PMM		PMM (MI) ⟷	PMM (MNAR)

Type(i), Type(ii)の仮定と感度パラメータ

欠測を含むデータの解析を行う際，モデルは部分的にtype(i), (ii)の二種類の仮定に分けられ，それぞれ感度分析の方法は異なる．

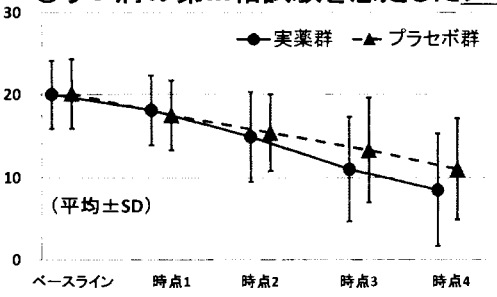
Type (ii) : **観測データの分布**に対する**検証可能**な仮定
→ 仮定の正しさ(モデルの適合度など)をみる．

Type (i) : **欠測データの分布**に対する**検証不可能**な仮定
→観測データの分布情報 ＋ 感度パラメータを用いることで，解析者が欠測データの分布を任意に仮定する．(MNARの仮定の下での解析が可能となる．)
→欠測メカニズムがMARの仮定から離れることに対する(MARを仮定した)主解析の結果の頑健性を確認する．

本セッションでは，NRC(2010)で紹介されている方法に準拠し，SMとPMMのモデル及び感度パラメータを用いた欠測メカニズム(MAR or MNAR)に対する感度分析の方法を紹介する．

解析対象データ

◎うつ病の第III相試験を想定したシミュレーションデータ



- ●主要評価項目：HAM-D
  → スコア低下：改善
  （解析には変化量使用）
- ●実薬群 vs プラセボ群
  ・1群100例（ベースライン時）

◎単調な欠測のみ

	ベースライン		時点1		時点2		時点3		時点4	
	例数	平均(SD)	例数	平均(SD)	例数	平均(SD)	例数	平均(SD)	例数	平均(SD)
実薬群	100	20.0 (4.1)	93	18.1 (4.2)	89	14.9 (5.4)	84	11.0 (6.3)	83	8.5 (6.8)
プラセボ群	100	20.1 (4.2)	90	17.5 (4.2)	87	15.4 (4.6)	85	13.3 (6.3)	80	11.0 (6.1)

11

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

SMを用いた感度分析

$$f(\mathbf{Y}_i, \mathbf{R}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{Y}_i|\boldsymbol{\theta}) \cdot f(\mathbf{R}_i|\mathbf{Y}_i, \boldsymbol{\psi})$$
$$= f(\mathbf{Y}_i^o, \mathbf{Y}_i^m|\boldsymbol{\theta}) \cdot \underbrace{f(\mathbf{R}_i|\mathbf{Y}_i^o, \mathbf{Y}_i^m, \boldsymbol{\psi})}_{\text{Type (i) の仮定}}$$

$$\text{logit}\{Pr(R_{ij} = 0|R_{i1} = 1, \cdots, R_{i,j-1} = 1, \mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\psi})\}$$
$$= \psi_0 + \psi_1 Y_{i,j-1} + \boxed{\psi_2} Y_{ij}$$

感度パラメータ：値を解析者が設定
色々な値を代入して結果の頑健性をみる

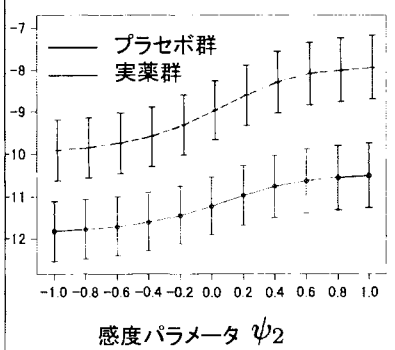
12

SMを用いた感度分析

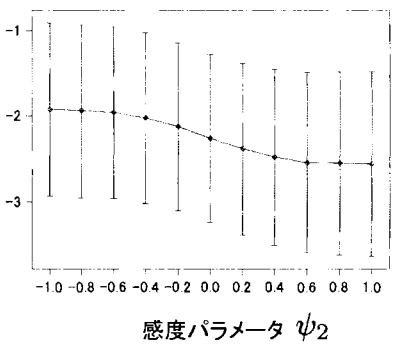
- SMでは，欠測モデル，応答変数モデル共にType(i) の仮定である.
- 応答変数について，群間差が評価できる範囲で，可能な限りモデルを仮定しないセミパラメトリックな方法もある.（IPW法：Inverse Probability Weightning）
- SMの感度パラメータは，現在測定された応答変数の大きさが欠測の有無に与える影響度合いを表している.
- $\psi_2 = 0$ のとき，MAR. それ以外の場合，MNARを意味している.

シミュレーションデータを用いた解析例 ～SMによる感度分析の結果～

- 各群の点推定値±SE



- 群間の差の点推定値±SE



PMMを用いた感度分析

$$f\left(Y_i^o, Y_i^m, R_i \setminus X_i, \beta\right)$$
$$= f\left(R_i \setminus X_i\right) f\left(Y_i^o \setminus R_i, X_i, \beta\right) \boxed{f\left(Y_i^m \setminus Y_i^o, R_i, X_i, \beta\right)}$$

未観測のデータのモデルに対する感度をみる

【目的】
制約条件を置くことに加えて，未観測のデータの応答変数の分布と観測データの応答変数の分布の乖離を評価することで欠測データのメカニズムMARの乖離を評価する.

PMMの経時測定モデル

今回の解析では，全ての未観測データに対して-Δを付加

$$f_t(y_1, \ldots. y_T) = f_t(y_1, \ldots. y_t)\, f_t(y_{t+1} | y_1, \ldots. y_t)\, f_t(y_{t+2}, \ldots, y_T | y_1, \ldots. y_{t+1})$$
$$= f_t(y_1, \ldots. y_t) \boxed{f_t(y_{t+1} | y_1, \ldots, y_t)} \prod_{s=t+2}^{T} \boxed{f_t(y_s | y_1, \ldots, y_{s-1})}$$
$$s \geq t+2$$

NFMVの制約条件
$s \geq t+2$に対して，

経時測定データにおける
未観測のデータの感度対象

$$\boxed{f(y_s | y_1, \ldots, y_{s-1}, r = t)} = f(y_s | y_1, \ldots, y_{s-1}, \boxed{r \geq s-1})$$

ただし，上記のr=s-1のときの条件付き分布と下記の分布

$$\boxed{f(y_{t+1} | y_1, \cdots y_t. r = t)}$$

は特定されていない．よって，追加のNCMV(+Δ)の制約条件を付加

$$\boxed{\begin{array}{l} f(y_s \setminus y_1, .., y_{s-1}, r = s-1) = f(y_s \setminus y_1, .., y_{s-1}, r = s) \\ f(y_{t+1} \setminus y_1, .., y_t, r = t) = f(y_{t+1} \setminus y_1, .., y_t, r = t+1) \end{array}}$$

OR

$$\boxed{\begin{array}{l} f(y_s - \Delta \setminus y_1, .., y_{s-1}, r = s) \\ f(y_{t+1} - \Delta \setminus y_1, .., y_t, r = t+1) \end{array}}$$

2時点（シンプル）の場合

興味あるパラメータ:Yの平均μ

未観測の人（R=0）の平均と観測されている人（R=1）の平均の乖離を感度パラメータΔで単純に表現

$\mu_0 = \mu_1 + \Delta \Leftrightarrow E(Y/R=0) = E(Y/R=1) + \Delta$

一般的な場合 $g(\mu_0) = g(\mu_1) + \Delta \Leftrightarrow E(Y/R=0) = g^{-1}(g(E(Y/R=1)) + \Delta)$

解析担当者が関数gを指定，単純ケースでは $g(\mu) = \mu$

検証不可能な仮定

Yの平均μは下記のように表現

$\mu = \Pr(R=1) \times \mu_1 + \Pr(R=0) \times \mu_0 = \pi\mu_1 + (1-\pi) g^{-1}(g(\mu_1) + \Delta)$

$\forall a \in R$に対して$g(a) = a$の場合，$\mu = \pi\mu_1 + (1-\pi)(\mu_1 + \Delta)$

17

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

- ある範囲の感度パラメータΔに対して，MNARのもとでの感度解析の一つとなる．

$$\begin{cases} \Delta = 0 \Leftrightarrow \mu = \mu_1 \Leftrightarrow MAR \\ \Delta \neq 0 \Leftrightarrow \mu \neq \mu_1 \Leftrightarrow MNAR \end{cases}$$

欠測メカニズムMARの仮定から乖離した μ にどの程度影響が出るか調査

【PMMの枠組み】

- 観測されている同じ測定時点のデータY_0をもち，未観測の人と観測されている人のY_1の分布を特定

$$g(E(Y_1 \setminus Y_0, R=0)) = g(E(Y_1 \setminus Y_0, R=1)) + \Delta$$

関数g及びΔを含めたモデルを規定が必要．ここでは単純な回帰モデルを示す．

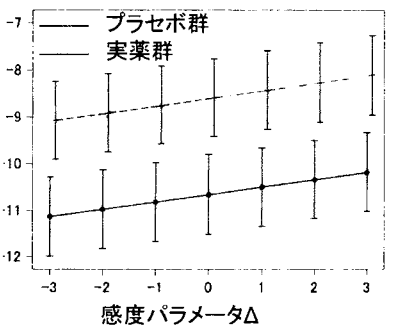
18

ベースライン調整

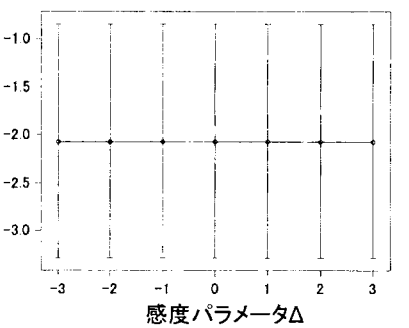
- ここでは単純な回帰モデルの場合を示す．
  $E(Y_1 \setminus Y_0, R=1) = \beta_0 + \beta_1 Y_0$ のモデルを仮定すると，
  $E(Y_1 \setminus Y_0, R=0) = E(Y_1 \setminus Y_0, R=1) + \Delta = \beta_0 + \beta_1 Y_0 + \boxed{\Delta}$
- 欠測データの分布の平均は回帰予測の標本平均により推定
- 得られた$\hat{\beta}_0, \hat{\beta}_1$はR=1において，$Y_0$を与えたうえでの$Y_1$への回帰により得られた推定値

シミュレーションデータを用いた解析例
～PMM（NFMV-NCMV＋Δ）による感度分析の結果～

□ 各群の点推定値±SE



□ 群間の差の点推定値±SE



SMとPMMの感度分析の違い

SM

$$f(\mathbf{Y}_i, \mathbf{R}_i|\theta, \psi) = f(\mathbf{Y}_i^o, \mathbf{Y}_i^m|\theta) \cdot \boxed{f(\mathbf{R}_i|\mathbf{Y}_i^o, \mathbf{Y}_i^m, \psi)}$$

脱落確率のモデルに対する感度をみる

PMM

$$f(\mathbf{Y}_i, \mathbf{R}_i|\theta, \psi) = f(\mathbf{R}_i|\psi) \cdot f(\mathbf{Y}_i^o|\mathbf{R}_i, \theta) \cdot \boxed{f(\mathbf{Y}_i^m|\mathbf{Y}_i^o, \mathbf{R}_i, \theta)}$$

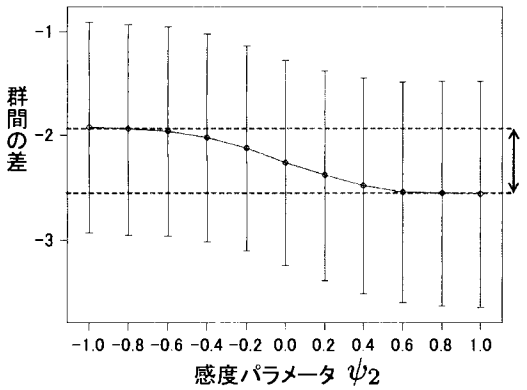
欠測データのモデルに対する感度をみる

本感度分析の結果の考察方法

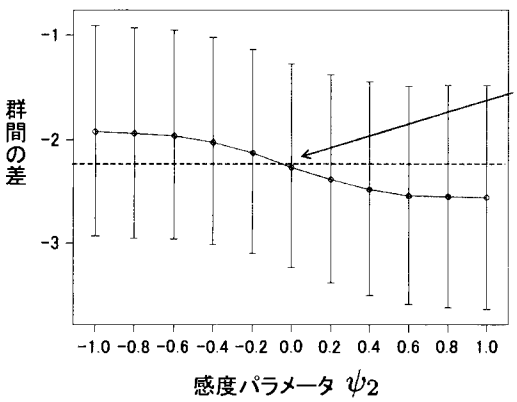
□欠測メカニズムに対する感度分析を含む解析手順の提案(前スライドより)
1. 主解析はMARを仮定した解析
2. 感度分析として，MNARの仮定の下で解析
3. MARを仮定した解析の結果の頑健性を確認し，試験結果の妥当性を述べる.

□ NRC (2010)では3種類の頑健性の確認方法が提案されている.（あくまで一案であり,今後も研究が必要）

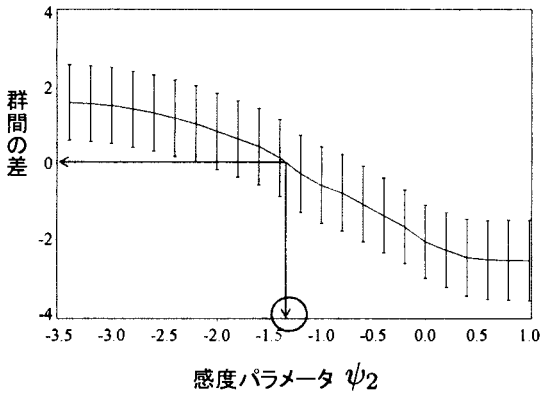
①感度パラメータに関して妥当な範囲(下限値～上限値)を指定し，この範囲内で点推定や95%信頼区間を算出
(ここでの95%信頼区間は,標本のばらつき+モデルの不確実性におけるばらつきを意味しているため，95% Confidence Region と呼ばれることもある.)



②感度パラメータに関して妥当な範囲(下限値～上限値)を指定し，複数の点推定値に対して平均やSEを求め，一つの結果にまとめる.



③MARに基づく推測を行い，MARの結果が変わるような感度パラメータの値を特定すること．もし，この値が妥当な範囲に含まれているならば，MARの結果が疑わしい可能性がある．



群間の差が0になる感度パラメータψ_2の値が科学的に妥当な範囲であるか考察する

☆NRC(2010)では，以下のコメントも残している

主解析及び感度分析の解析結果のどちらを重視すべきか？

- 極端な仮定をおいた場合の感度分析の結果はあまり重視されないが，主解析での仮定と同等の範囲内で行われた感度分析の結果は重視される．

- 主解析の結果が感度分析の結果と反対になることがあったとしても，そのときの感度分析の仮定が極端なものであるならば主解析の結果を支持するのは合理的

=>極度な仮定ではなく，妥当な仮定の範囲で主解析の結果が支持されていることが重要

参考文献

- National Research Council (NRC). The Prevention and Treatment of Missing Data in Clinical Trials. Washington, DC: The National Academies Press, 2010
- Craig H. Mallinckrodt (2013), Preventing and Treating Missing Data in Longitudinal Clinical Trials: A Practical Guide, Cambridge University Press
- Bohdana Ratitch, Michael O'Kelly, and Robert Tosiello, Missing data in clinical trials:from clinical assumptions to statistical analysis using pattern mixture models, Pharmaceutical. Statistics. 2013, 12, 337-347.
- SAS macro. missingdata.org.uk . http://missingdata.lshtm.ac.uk/index.php?view=category&id=61%3Amnar-methods&option=com_content&Itemid=137

【企画セッション】欠測のあるデータに対する各種解析手法と欠測メカニズムに対する感度分析

(5) まとめと質疑応答

土居 正明[1)]
日本製薬工業協会 医薬品評価委員会 データサイエンス部会 タスクフォース4
欠測のあるデータに対する解析方法論・SASプログラム検討チーム

1)東レ株式会社

Summary of the session and Q&A

Masaaki Doi[1)]
The team for statistical methodologies and SAS programming of data analysis with missing data, task force 4, data science expert committee, drug evaluation committee, Japan Pharmaceutical Manufacturers Association.

1) Toray industries, Inc.

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨：

本セッションの内容をまとめ，質疑応答を行う．

キーワード：SM, MMRM, PI, PMM, SPM, 感度分析，
感度パラメータ

本セッションの目標

欠測のあるデータの解析に対する

① 考え方の理解

② SASでの実行方法の理解

(a) SM, MMRM, MI, PMM, SPM
(b) 欠測メカニズムに対する感度分析

本セッションのまとめ

◎MARを仮定する方法
・MMRM, MI, PMM (ACMV)　※単調な欠測を仮定

◎MNARを仮定する方法
・SM, PMM (ACMV以外), SPM

主解析ではMARを仮定することが多い

がMARかMNARかは、データからは確定できない

→ **感度分析**が必要

＜欠測メカニズムに対する感度分析＞

・Type (i) の仮定 → データから仮定の妥当性が確認できない
→ 色々なモデルを当てはめて頑健性をみる

・Type (ii)の仮定 → データから仮定の妥当性が確認できる
→ モデル診断など

◎MARかMNARか？はType (i)の仮定

→ 感度パラメータを用いたSM, PMM

検討できていない点1

・その他の解析手法

→ wGEE, Bayes, IPW, AIPW (Doubly Robust)...etc.

・感度分析の結果の要約方法

→ 主解析と感度分析で結果がある程度以上異なる場合など，一般的な状況でコンセンサスのとれた要約方法はまだない．

→ 統計家の役割が大きい．

検討できていない点2

・主解析の検討方法
・欠測メカニズム以外に対する感度分析

→ ・Estimandの選択？
・完全データの分布？
・外れ値の検討？

結局，解析全体として何をすれば？

→ 続きは「欠測のあるデータに対する総合的な感度分析と主解析の選択」で．

欠測のあるデータに対する総合的な感度分析と主解析の選択

土居正明[1)2)]，大浦智紀[1)3)]，大江基貴[1)4)]，駒嵜弘[1)5)]，髙橋文博[1)6)]，
縄田成毅[1)7)]，藤原正和[1)8)]，横溝孝明[1)9)]，横山雄一[1)10)]
1) 日本製薬工業協会 医薬品評価委員会　データサイエンス部会 タスクフォース 4
欠測のあるデータに対する解析方法論・SAS プログラム検討チーム
2)東レ株式会社　3)日本イーライリリー株式会社 4)株式会社大塚製薬工場
5)マルホ製薬株式会社　6)田辺三菱製薬株式会社 7)杏林製薬株式会社
8)塩野義製薬株式会社　9)大正製薬株式会社 10)持田製薬株式会社

Comprehensive sensitivity analyses and choice of primary analysis when some data are missing.

Masaaki Doi[1)2)], Tomonori Oura[1)3)], Motoki Oe[1)4)], Hiroshi Komazaki[1)5)], Fumihiro Takahashi[1)6)],
Sigeki Nawata[1)7)], Masakazu Fujiwara[1)8)], Takaaki Yokomizo[1)9)], Yuichi Yokoyama[1)10)]
1) The team for statistical methodologies and SAS programming of data analysis with missing data, task force 4,
data science expert committee, drug evaluation committee, Japan Pharmaceutical Manufacturers Association,
2) Toray Industries, Inc. 3) Eli Lilly Japan K.K. 4) Otsuka Pharmaceutical factory, Inc. 5) Maruho Co, Ltd.
6) Mitsubishi Tanabe Pharma Corp. 7)Kyorin Pharmaceutical Co., Ltd. 8) Shionogi & Co., Ltd.
9) Taisho Pharmaceutical Co., Ltd. 10) Mochida Pharmaceutical Co., Ltd.

要旨

欠測のあるデータに対する，主解析・感度分析を含めた解析の全体像を Mallinckrodt (2013)をもとに検討する．各解析手法は，応答変数や欠測メカニズムに対する様々な仮定のもとで妥当性が保証されている．そのため，主解析の結果の妥当性を示すためには (i) 他の解析を行っても結果が変わらない，(ii) 仮定の妥当性を確認する，等の感度分析が必要となる．欠測のあるデータに対する感度分析の方法としては NRC (2010)で検討はされているものの，欠測メカニズムに対する感度分析に限定されており，主解析・感度分析を含めた解析方法を選択する上での具体的な指針とはなりにくい．そこで本稿では，NRC (2010)を発展させた Mallincrkodt (2013)に従い，estimand 等も視野に入れた総合的な解析の検討を行う．また，MAR を仮定した場合，主解析として複数の解析方法が考えられる．計画段階で主解析を選択する際の，シミュレーションによる性能評価も行う．

キーワード：感度分析，Estimand，MAR，MNAR，Type (i) の仮定，Type(ii)の仮定，MMRM，MI，wGEE，SM，PMM，SPM，pMI

1. はじめに

欠測のあるデータに対して，欠測を考慮した解析が必要であることは広く認識されてきている．その結果，個別の手法に対する知識は広まってきたが，「主解析を選択する際に何が重要であるか」「どのような感度

分析が必要か」等の基本的な問いに対しては，一般に合意された考え方は存在しなかった．そのような中，NRC (2010)では，主解析・感度分析に対する考え方が提案されたが，扱われている感度分析が限定的であること，解析の全体像が見えにくいことなどから，実際に解析を実施する際の十分なガイドラインとは言い難かった．その後提案された Mallinckrodt (2013)では，主解析・感度分析を含めた解析の全体像が示された．そこで本稿では主に Mallickrodt (2013)に従い，欠測のあるデータの解析の全体像を示すことを目的とする．

なお，本稿では単調な欠測のみを扱うこととする．また，解析プログラムとして Missingdata.org.uk (www.missingdata.org.uk) にて公開されているマクロプログラムを適宜用いた．

2. 感度分析と解析の全体像

2.1 NRC (2010)の感度分析

NRC (2010)では，感度分析の種類として(i)完全データの分布に対する感度分析，(ii)外れ値・外れた症例に対する感度分析，(iii)欠測メカニズムに対する感度分析，が挙げられている．そして，最も重要なものは(iii)欠測メカニズムに対する感度分析である，として，(iii)に対してのみ具体的な方法の提案がなされた．

2.2 Mallinckrodt (2013)の感度分析と解析の全体像（Analytic Road Map）

Mallinckrodt (2013)は NRC (2010)同様，欠測メカニズムに対する感度分析が最も重要と指摘した上で，NRC (2010)では扱われていない感度分析についても十分に記載し，主解析・感度分析を含めた解析の全体像を示した"Analytic Road Map"を提示した．図 1 は Analytic Road Map を微修正したものである．

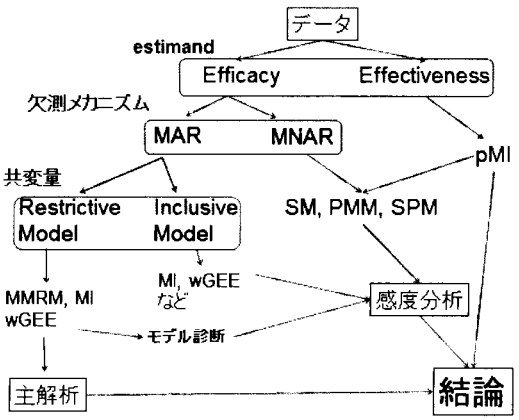


図 1. Analytic Road Map（著者らによる加工あり）

本稿では，図 1 をもとに，主要な解析と感度分析を含めた解析全体に対する検討を行う．特に，estimand，欠測メカニズム，完全データの分布（共変量・モデル診断含む）に注目する．

2.3 Type (i)・Type (ii)の仮定とそれぞれに対する感度分析

欠測のあるデータの解析に限らず，各統計手法は一定の仮定が正しい場合の正当性が示されている．そのため，得られたデータが仮定を満たしているかどうかを確認するのは極めて重要である．一方，特に欠測のあるデータを扱う場合，「データが欠測しているため，仮定の正しさを検討することができない」という状態も生じうる．このような状況を受け，NRC (2010)では，仮定を以下の 2 通りに分けた．

- Type (i)：検証不能(untestable)な仮定
- Type (ii)：検証可能(testable)な仮定

それぞれに対して，感度分析は以下のように定義される．

・Type (i)：色々なモデルを当てはめ，パラメータ推定値が安定した値となっているかどうかを検討する

・Type (ii)：モデル適合をみる統計量を用いて，仮定が妥当であるかを検討する

つまり Type (i) の仮定に対しては「仮定の正しさ」は考えず，「結果の頑健性」の担保に専念するのである．

2.4 何に対する感度分析か？

たとえば主解析として MAR を仮定した MMRM を選択し，感度分析として同じく MAR を仮定した MI, wGEE を行ったとする（MMRM という用語の使い方は，大江ら(2014)と同様とする）．この場合，各解析で仮定している分布の違い等に対する感度分析は行えるが，NRC (2010)や Mallinckrodt (2013)で「最も重要」と指摘されている，欠測メカニズムに対する感度分析は行えていない．そのため，実際の欠測メカニズムが MNAR であった場合，感度分析と合わせても楽観的な結論となる可能性が否定できない．このように，感度分析を行う際は「主解析でどのような仮定をしているか」を把握した上で，「どの仮定に対する感度を検討しているか」に注目することが重要である．本稿では特に「Estimand」「欠測メカニズム」「完全データの分布」に対する感度分析を検討する．

3.Estimand に対する感度分析

欠測のあるデータに対しては，様々な解析手法がある．各手法が「本当にその試験で知りたいものを推定できているか」を考えるためにも，試験の目的をより明確化することが望ましい．そのために estimand を考えることが役に立つ．Estimand とは，一般には"what is being estimated"のことであり，「推定の対象」程度の非常に広い意味をもつ．Mallinckrodt (2013)は経時データを扱う試験の estimand の要素として「パラメータ（例．平均の群間差）」「時点または曝露期間（例．投与期間 8 週目）」「アウトカム（例．拡張期血圧）」「対象となる集団（例．高血圧と診断された患者）」「中止後に治療(rescue medication)が行われた場合，その後に得られたデータは解析に含めるかどうか」等がある，と述べている．estimand の種類として，NRC (2010)では 5 種類が提案され，これに 1 つ追加した 6 種類が Mallinckrodt et al. (2012)や Mallinckrodt (2013)で提案されている．また，Mallinckrodt et al. (2014)には，その中の 3 種類が記載されている．

3.1 efficacy と effectiveness

Estimand について考える際，efficacy と effectiveness の区別を認識しておくことが重要である．Mallinckrodt et al. (2012)によると，efficacy とは，計画通りに投与された場合の薬剤の影響であり，per-protocol estimand とも呼ばれる．一方，effectiveness とは，実際に投与された薬剤の影響であり，ITT estimand とも呼ばれる．

3.2 Mallinckrodt et al. (2012)，Mallinckrodt (2013)の 6 種類の estimand

以下，Mallinckrodt et al. (2012)で，慢性疾患の臨床試験を想定しつつ，精神疾患，痛み，糖尿病等にも使用可能なものとして提案された 6 種類の estimand について述べる．

Estimand 1：全てのランダム化された症例に対する，計画された時点でのアウトカムの改善具合の差

Estimand 2：最初の治療に耐えられた症例のアウトカムの改善具合の差

Estimand 3：全症例が治療を完了できたと仮定した場合のアウトカムの改善具合の差

Estimand 4：治療を継続できた期間の Area Under the outcome Curve の差

Estimand 5：治療を継続できた期間におけるアウトカムの改善具合の差

Estimand 6：全てのランダム化された症例に対する，計画された時点での，最初に割り付けられた治療によるアウトカムの改善具合の差

表 1. 6 種類の estimand（Mallinckrodt et al., 2012; Mallinckrodt, 2013）

Estimand	仮説	推測の対象	被験者	時点	Rescue Medication 後のデータ
1	Effectiveness	割り付け群	全被験者	計画された時点	主解析に含める
2	Efficacy	最初に割り付けられた治療	最初に耐えられた被験者のみ	計画された時点	主解析に含めない
3	Efficacy	最初に割り付けられた治療	全被験者	計画された時点	主解析に含めない
4	Effectiveness	最初に割り付けられた治療	全被験者	未定義	主解析に含めない
5	Effectiveness	最初に割り付けられた治療	全被験者	未定義	主解析に含めない
6	Effectiveness	最初に割り付けられた治療	全被験者	計画された時点	補完することが望ましい

各 estimand の特徴は以下の通りである.

Estimand 1 は，実際に投与された薬剤ではなく，割り付け群の影響をみる．中止後に rescue medication が行われ，その後にデータが得られた場合，そのデータも解析に利用する．しかし，(1)多くの場合，臨床試験で興味があるのは最初に割り付けられた薬剤の効果であること，(2)最初に割り付けられた薬剤の効果が rescue medication によって過大評価もしくは過小評価されること，などから Mallinckrodt et al. (2012)では今回の状況では estimand 1 は主要な estimand として適切ではない，と指摘されている.

Estimand 2 はランダム化前の run-in 期間に全員に実薬を投与し，治療に耐えられ，継続できた被験者にのみランダム化を行い，治療効果をみる．中止症例が減るため，efficacy の評価に役立つ．一方で注意すべき点には，(1)この estimand で評価できるのは全患者集団のうち初期の治療に耐えられる部分集団に対してであるため，一般化可能性に疑問が残ること，(2)実際の治療の際にはどの患者が初期の治療に耐えられるか分からないこと，(3)ランダム化前の run-in 期間を設定していなければ本 estimand を検討できないこと，などがある.

Estimand 3 は，全症例が治療を完了したと仮定した場合の群間差を評価するものである．これは理想的な状況であるため，中止症例がどの程度有効性に影響を与えるかを別途評価することが必要である．具体的には，この estimand を主解析とした場合，effectiveness をみる感度分析を行うべきである.

Estimand 4, 5 は最初に割り付けられた薬剤の effectivness を評価するもので，アウトカムの大きさと治療に耐えられた期間を合わせて数値化する．従って，中止によるデータの欠測は生じない．一方，中止後も治療効果が持続する場合でなければ，計画された評価時点での effectiveness を過大評価する傾向にある.

Estimand 6 は estimand 1 と似ているが，中止後に rescue medication を行った後のデータの取り扱いが異なる．Estimand 1 はそのままデータを使用したが，estimand 6 ではこのデータにはバイアスが入っているものとして，補完した値を解析に用いるのが望ましいとされる．これは，興味の対象が「計画された評価時点での，最初に割り付けられた治療の効果」であるからである．そのため，中止後に rescue medication が行われた場合は，

その影響を取りのぞき，中止後は無治療であった場合の推定を行う．なお，補完の方法として，Mallinckrodt et al. (2012)では pMI (placebo Multiple Imputation)が推奨されている．pMI はプラセボ群の推移をもとに，中止後のデータの予測分布を構成した上で，Multiple Imputation を行う方法である．詳細は Mallinckrodt et al. (2012), Mallinckrodt (2013)等を参照．なお，NRC (2010)では estimand 1～5 が提案され，Mallinckrodt et al. (2014)では，estimand 1 が estimand A，estimand 3 が estimand B，estimand 6 が estimand C と呼ばれている．

3.3 主解析に対する Estimand と Estimand に対する感度分析

Estimand は，疾患領域・薬剤の特性・治験の状況等を考慮して，試験毎に設定すべきである．なお，異なる estimand に対する解析は目的が異なっているため，厳密には感度分析と考えるべきではないかもしれない．しかし，結果を比較した上で解釈する，という観点から，本稿では感度分析に含める．

以下，一例として Mallinckrodt et al. (2012)で検討された慢性疾患の第 III 相試験の estimand について述べる．まず，主要な estimand は estimand 3 とした．これは，興味の対象が最初に割り付けられた薬剤の efficacy であるからである．ただし，estimand 3 は上で述べた通り全症例が治療を完了したと仮定した場合の治療効果である．そのため，欠測症例の影響を評価するための感度分析として effectiveness を検討することが推奨されている．これより，estimand 6 の検討も行った．解析方法としては，estimand 3 に対応する主解析としては MMRM，estimand 6 に対する解析としては pMI が用いられた．また，今回の状況は慢性疾患であるため，中止後に治療効果が弱まることが想定された．従って，estimand 4, 5 は望ましくないと考えられた．

4. 欠測メカニズムに対する感度分析

一般に，主解析の際は欠測メカニズムとして MAR を仮定することが多い．しかし，これは MAR が多くの臨床試験で実際に成り立っている，ということを必ずしも意味するものではない．Mallinckrodt et al. (2008)によると「計画段階で MAR となるように十分に計画を立てておくべきである」「MNAR を仮定した解析は，仮定が間違っていた場合に MAR を仮定した解析よりも大きなバイアスが入りやすい」などを考慮した上での取り扱いであり，欠測メカニズムが MNAR である可能性は否定されていない．また，欠測メカニズムが MAR であるか MNAR であるかは Type (i)の仮定であり，観測データからは検証できない．そのため，MNAR である可能性は常に考慮しておくべきである．

以上より，主解析に MAR を仮定した解析を用いた場合，感度分析として MNAR を仮定した解析を行うことは大変重要である．先にも述べた通り，NRC (2010)や Mallinckrodt (2013)では，感度分析全体の中で欠測メカニズムに対する感度分析が最も重要であると指摘されている．

具体的な解析方法の詳細は大江ら(2014)，高橋ら(2014)，駒嵜ら(2014)参照．

5. 完全データの分布に対する感度分析

5.1 完全データの分布

Mallinckrodt (2013)に従い，完全データに対して，共変量の検討・残差診断・影響診断・分散共分散構造の検討の 4 項目を考える．

【共変量の検討】

まず，NRC (2010)に従い，共変量を 2 種類に分ける．

・デザイン変数：投与群やベースライン時に得られる共変量で，全ての症例に対して観測され，主解析に用いられる共変量.

・補助変数　　：欠測したデータに対する推測に利用できる変数．投与前のものもあれば投与後のものもある．コンプライアンスや副作用など，主解析の共変量には用いられないが脱落確率や欠測データの分布のモデリングに役立つ共変量.

ここで,デザイン変数のみを含めたモデルを Restrictive Model, 補助変数まで含めたモデルを Inclusive Model と呼ぶ．主解析には Restrictive Model を用いた解析を行い，適当な補助変数がある場合は，感度分析として Inclusive Model を用いた解析を行うことが考えられる．なお，Mallinckrodt (2013)は，ランダム化後に得られ，薬効と交絡する変数であっても,欠測確率の予測に用いる補助変数とすることができるものは存在しうる（ただし，このような補助変数は応答変数のモデルには含めるべきではない），と述べている.

【残差診断】

モデル適合度をみる方法として，残差診断はよく用いられる方法である．主解析として MMRM を用いる場合，SAS の MIXED PROCEDURE のモデルステートメントの RESIDUAL オプションで様々な残差を出力できる．算出された残差に対して，(1)全体的なプロットを眺めて傾向をみる，(2)閾値を決め，閾値を超えるものを外れ値と考え，それを除外した上で主解析を再度行うことにより，外れ値の推定値に与える影響を評価する，などが考えられる．なお，Mallinckrodt (2013)では Student 化残差の絶対値が 2 以上の場合を外れ値と考え，除外した解析との結果を比較している.

【影響診断】

次に，影響の大きい症例や施設など，影響の大きいクラスタの探索方法について述べる．以下，主解析として MMRM を想定する．容易に実行できるのは，Cook の D 統計量を用いる方法である．これは，MIXED PROCEDURE のモデルステートメントの INFLUENCE オプションで算出できる．Cook の D 統計量の値が大きい症例や施設を影響の大きい症例・施設と考え，それを除外した上で主解析を再度行うことにより，影響の大きい症例や施設の，主解析の結果に与える影響が検討できる.

別の方法として，local influence を用いる方法があるが，現段階では実行が難しいため，本稿ではこれ以上触れない．詳細は Verbeke and Molenberghs (2000)，Molenberghs and Kenward (2007)などを参照.

【分散共分散構造の検討】

分散共分散構造の違いが結果に与える影響についても，感度分析の対象となりうる．MMRM では，分散共分散構造として,仮定が少なくあてはまりのよいことが想定される無構造(unstructured)を指定する場合が多いが，これ以外の分散共分散構造を用いた場合の結果と比較することで，結果の頑健性を確認することができる．また，AIC 等のモデル評価基準を用いて妥当な相関構造について検討することもできる.

6. 主解析・感度分析の実行

以上で述べた方法に従い，図 1 の"Analytic Road Map"をもとに，主解析・感度分析を検討した.

6.1 試験計画・解析手法

うつ病の第 III 相試験を想定し，以下の通りの解析を設定した．データは 6.2 解析対象のデータで示す..

＜estimand に対する計画段階での検討＞

本試験はうつ病（精神疾患）の第 III 相試験を想定した．応答変数としては HAM-D スコアを用い，時点 4（8 週目）の平均の差を評価することを考えた．さらに，3.3 主解析に対する estimand と estimand に対する感

度分析 に従い，efficacy の評価を目標とし，主要な estimand を estimand 3 とした．また，中止症例の影響を評価する感度分析として，estimand 6 を用いた effectiveness の評価も実施することとした．

＜主解析＞

MMRM とした．詳細は以下の通りである．

・共変量：（連続値）ベースライン値，（カテゴリ値）投与群，時点，投与群と時点の交互作用

・変量効果：被験者の影響を誤差と合わせてモデル化するため，明示的には特定しない

・相関構造：Unstructured（被験者ごと）

収束しなかった場合①初期値を Fisher's Scoring 法で得られた値とする

②相関構造を Toeplitz, Heterogeneous CS，AR(1)，CS，VC の順に指定する．

・推定方法：REML

・自由度調整方法：Kenward Rodger

・帰無仮説：時点 4 で群間差が 0

・有意水準：両側 5%

＜感度分析 1：モデル適合の検討＞

主解析に対する外れ値や影響の大きい症例の検討のため，残差診断と影響診断を行った．

・残差診断：Student 化残差に対して①全体的な傾向を観察した，②絶対値が 2 以上の場合に外れ値とみなし，除外した解析を実施した．

・影響診断 ：Cook の D 統計量が 0.03 以上の症例を影響の強い症例とみなし，除外した解析を実施した．

＜感度分析 2：欠測メカニズムに対する感度分析＞

欠測メカニズムに対する感度を検討するため，以下の各モデルに対する解析を実施した．解析方法の詳細は大江ら(2014)，高橋ら(2014)，駒嵜ら(2014)参照．

・選択モデル：MNAR を仮定し，感度パラメータを-1～1 に対して 0.2 区切りで設定

・パターン混合モデル：NFMV (NCMV)を仮定し，感度パラメータを-3～3 に対して 1 区切りで設定

・共有パラメータモデル：変量切片を考慮したモデル

＜感度分析 3：estimand に対する感度分析＞

Effectiveness に対する検討として，placebo Multiple Imputation (pMI)を用いた解析を実施することとした．補完後のデータは MMRM で解析を行った．

以上の解析をまとめたものを以下の表 2 に示す．

表 2 主解析・感度分析の一覧

	仮説	Estimand	欠測メカニズム	解析手法	その他
主解析	Efficacy	3	MAR	MMRM	
感度分析 1	Efficacy		MAR	MMRM	モデル適合の検討
感度分析 2	Efficacy		MNAR	SM	感度パラメータ-1～1
			MNAR	PMM	NFMV (NCMV) 感度パラメータ-3～3
			MNAR	SPM	
感度分析 3	Effectiveness	6	MNAR	pMI	

6.2 解析対象のデータ

図 2，表 3 に示すシミュレーションデータを用いた．以下，応答変数の値の要約を示すが，解析にはベースラインからの変化量を用いた．なお，脱落確率は両群でほぼ等しく，時点 4 で 20%程度であった．

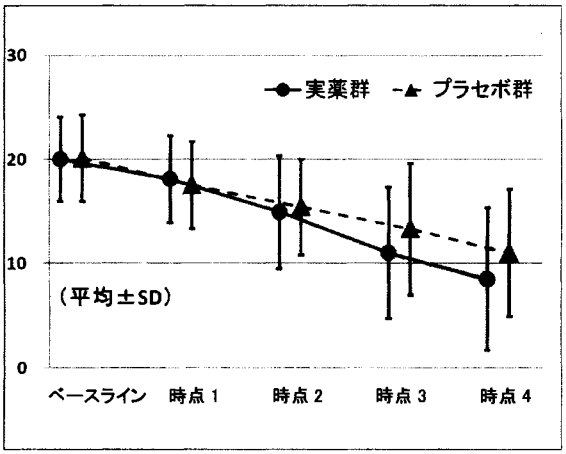


図 2　群ごとの HAM-D スコアの推移（平均±SD）

表 3　時点ごとの症例数・HAM-D スコアの平均・SD

	ベースライン		時点 1		時点 2		時点 3		時点 4	
	例数	平均 (SD)	例数	平均 (SD)	例数	平均 (SD)	例数	平均 (SD)	例数	平均 (SD)
実薬群	100	20.0 (4.1)	93	18.1 (4.2)	89	14.9 (5.4)	84	11.0 (6.3)	83	8.5 (6.8)
プラセボ群	100	20.1 (4.2)	90	17.5 (4.2)	87	15.4 (4.6)	85	13.3 (6.3)	80	11.0 (6.1)

6.3 解析結果

解析結果は，以下の通りである．

＜主解析＞

主解析である MMRM の結果を表 4 に示した．

表 4 主解析の結果

解析手法	時点 4 における各群の点推定値(SE)		群間差	群間差の SE	p 値
	実薬群	プラセボ群			
MMRM	-11.22 (0.69)	-8.97 (0.70)	-2.26	0.99	0.024

有意水準両側 5%で有意であり，群間差は-2.26 と十分に大きかったため，主解析としては有効性を示す結果となった．以下，感度分析でこの結果の頑健性を検討する．

＜感度分析 1：モデル適合の検討＞

Student 化残差の残差プロット，Cook の D 統計量のプロットを図 3, 4 に示した．図 3 より，やや負の値が

多く見えるものの，全体的に 0 に対してほぼ対称に分布しており，特に偏りはみられなかった．また，Cook の D 統計量が 0.03 を超える症例は 1 症例，Student 化残差の絶対値が 2 を超えるデータは 31 ポイント存在した．これらの症例・ポイントを除いて，主解析と同じモデルで MMRM による解析を実施した．主解析と比較した結果を表 5 に示した．群間差や群間差の SE はやや異なるものの，有効性を大きく減少させる傾向はみられなかった．

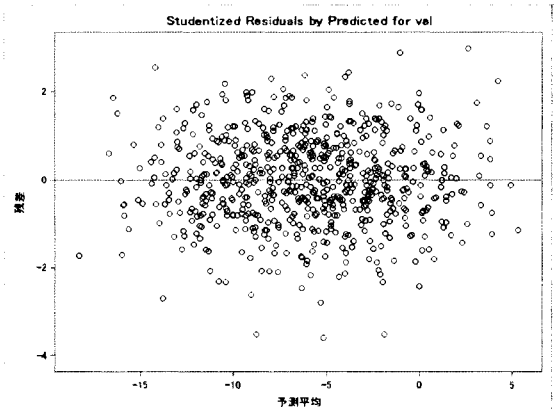


図 3 残差プロット（Student 化残差）

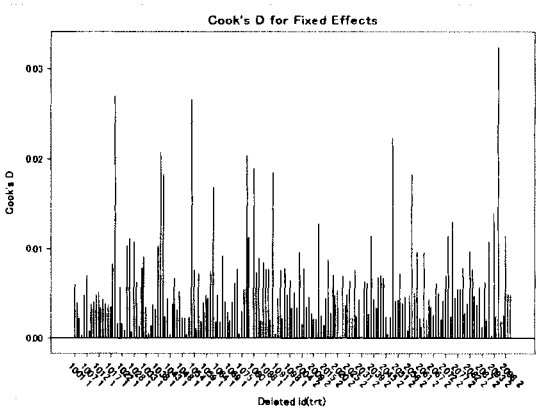


図 4 Cook の D 統計量

表 5　外れ値・影響の強い症例の検討

	除外判定基準	症例数	データ数	群間差	群間差の SE
主解析		183	691	-2.26	0.99
感度分析 1 （外れ値除外）	Student 化残差 絶対値 2 以上	183	660	-2.37	0.87
感度分析 1 （外れた症例除外）	Cook's D 0.03 以上	182	687	-2.24	1.00

＜感度分析 2：欠測メカニズムに対する感度分析＞

次に，欠測メカニズムに対する感度分析の結果をみる．まず，感度パラメータを含めた SM による解析結果を図 5, 6，表 6 に示した．図 5，6 の横軸は感度パラメータである．

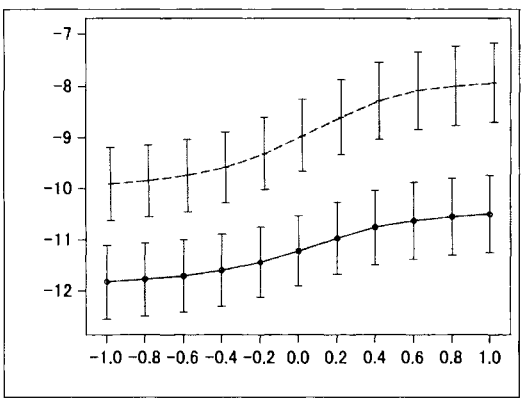


図 5. 感度パラメータごとの両群の点推定値±SE (SM)

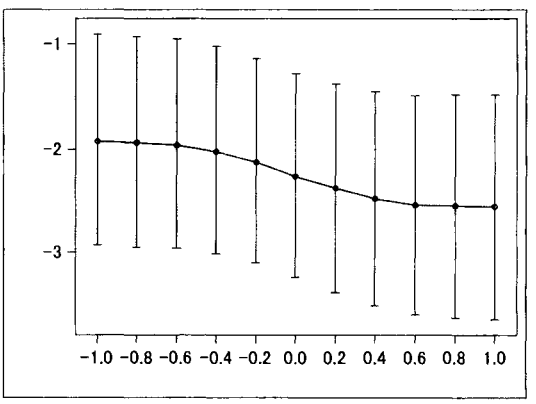


図 6. 感度パラメータごとの群間差±SE (SM)

感度パラメータの値が小さくなるほど，群間差の値が小さくなる傾向がみられた．群間差の最小値は-1.92であった．

表 6. 感度パラメータごとの解析結果(SM)

解析手法	感度パラメータ	時点 4 における各群の点推定値(SE)		群間差	群間差の SE
		実薬群	プラセボ群		
SM	-1.0	-11.83(0.72)	-9.91(0.72)	-1.92	1.01
	-0.8	-11.78(0.71)	-9.84(0.71)	-1.94	1.01
	-0.6	-11.71(0.70)	-9.74(0.71)	-1.96	1.00
	-0.4	-11.60(0.70)	-9.58(0.70)	-2.02	0.99
	-0.2	-11.44(0.69)	-9.31(0.70)	-2.12	0.98
	0.0	-11.22(0.69)	-8.96(0.70)	-2.26	0.98
	0.2	-10.98(0.71)	-8.60(0.72)	-2.38	1.00
	0.4	-10.76(0.73)	-8.29(0.74)	-2.48	1.03
	0.6	-10.63(0.75)	-8.09(0.75)	-2.54	1.05
	0.8	-10.55(0.76)	-8.00(0.76)	-2.55	1.07
	1.0	-10.50(0.76)	-7.94(0.76)	-2.56	1.08

次に，感度パラメータを用いた PMM の結果を図 7，8，表 7 に示した．

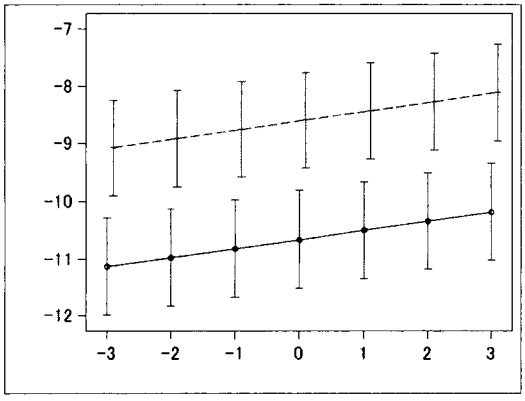


図 7. 感度パラメータごとの両群の点推定値±SE (PMM)

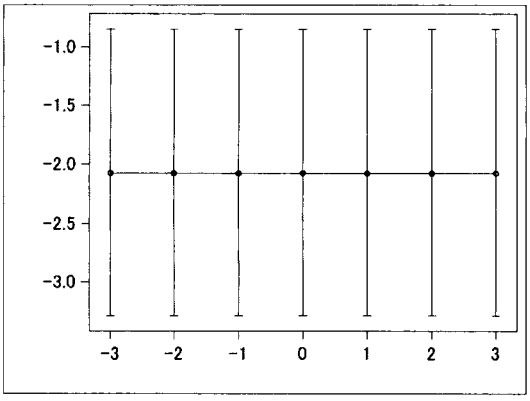


図 8. 感度パラメータごとの群間差±SE (PMM)

群間差の点推定値・SE は感度パラメータにほぼ影響を受けなかった．これは NRC (2010)の PMM を用いた感度分析の結果と同様の傾向である．

最後に，SPM の結果を表 8 に示した．本シミュレーションに用いたマクロでは，実薬群・プラセボ群の点推定値・SE は出力されなかった．群間差は-2.65 となり，主解析の MMRM よりやや大きくなった．

＜感度分析 3：estimand に対する感度分析＞

Effectiveness (estimand 6)を評価する解析として pMI を実行した結果を表 9 に示した．群間差は-2.06 となり，主解析の結果よりやや小さい値となった．プラセボ群のデータを元に中止後のデータを補完したため，主解析より保守的な値となった．

表 7. 感度パラメータごとの解析結果(PMM)

解析手法	感度パラメータ	時点 4 における各群の点推定値(SE)		群間差	群間差の SE
		実薬群	プラセボ群		
PMM : NFV (NCMV)	-3.0	-11.14(0.85)	-9.07(0.83)	-2.07	1.22
	-2.0	-10.98(0.85)	-8.91(0.83)	-2.07	1.22
	-1.0	-10.82(0.85)	-8.75(0.83)	-2.07	1.22
	0.0	-10.66(0.85)	-8.59(0.83)	-2.07	1.22
	1.0	-10.50(0.84)	-8.43(0.84)	-2.07	1.22
	2.0	-10.34(0.84)	-8.27(0.84)	-2.07	1.22
	3.0	-10.18(0.84)	-8.11(0.84)	-2.07	1.22

表 8. SPM の解析結果

解析手法	時点 4 における各群の点推定値(SE)		群間差	群間差の SE
	実薬群	プラセボ群		
SPM	-	-	-2.65	0.76

表 9. pMI の解析結果

解析手法	時点 4 における各群の点推定値(SE)		群間差	群間差の SE
	実薬群	プラセボ群		
pMI	-11.03 (0.70)	-8.97 (0.71)	-2.06	0.99

6.4 感度分析の結果のまとめ

以上，主解析を MMRM とし，(1)モデル適合，(2)感度パラメータ，(3)estimand に対する感度分析を実行した．検討した感度分析の範囲内では，主解析のモデル適合に問題点はみられなかった．次に，各解析の群間差の点推定値（の最大・最小）をまとめたものを表 10 に示した．

表 10. 主解析・感度分析の群間差一覧

	解析手法	その他	群間差		
				最小	最大
主解析	MMRM		-2.26		
感度分析 1	MMRM	モデル適合の検討		-2.24	-2.37
感度分析 2	SM	感度パラメータ-1～1		-1.92	-2.56
	PMM	NFMV (NCMV) 感度パラメータ-3～3		-2.07	-2.07
	SPM		-2.65		
感度分析 3	pMI		-2.06		

群間差は主解析と比較してやや小さくなるものもみられたが，点推定値は-1.92～2.65 であり，主解析の結果と大きく乖離するものではない考えられたため，本薬剤の有効性は安定していると考えられた．

6.5 実際に適用する際の注意

以上のような感度分析を実際に利用する際，以下の点などに注意が必要である．

① estimand を適切に選択する．

② 仮定を意識して適切な感度分析を計画・実行する．

③ 感度パラメータの適切な範囲を検討する．群ごとに異なるパラメータを用いることも検討する．

④ 公開されているマクロを使用する場合，適切にプログラムのバリデーションをとる．

⑤ 感度分析の結果が主解析の結果と比較的大きく異なる場合，解釈を慎重に行う．

特に，③（特に SM）⑤などは今後の課題と考えられている．

7. 主解析の選択

7.1 MAR を仮定した場合の主解析

次に，欠測メカニズムとして MAR を仮定した場合，計画段階でどの解析手法を主解析とするか，の選択のためにシミュレーションによる検討を行った．図 1 の主解析の候補である MMRM，wGEE，MI に LOCF（解析手法は共分散分析）を加えた 4 種類を比較した．各手法の詳細は大江ら(2014), 高橋ら(2014), NRC (2010), Mallinckrodt (2013)，Molenberghs and Kenward (2007)，O'Kelly and Ratitch (2014)等を参照せよ．

MMRM の詳細等，以下で特に触れない部分は，6.1 試験計画・解析手法と同様とする．その他の設定は以下の通りである．

【wGEE】脱落モデル：logistic モデル（共変量は主解析と同じ）

【MI】　補完モデル：投与群ごとの単調回帰モデル（共変量：ベースライン，各時点の変化量）．

補完回数：5 回

解析方法：共分散分析（共変量：投与群，ベースラインのみ）

【LOCF】解析方法：共分散分析（共変量：投与群，ベースラインのみ）

7.2 シミュレーションの設定

【完全データ】

以下のような状況を想定し，完全データを作成した．

・投与群：2 群　g（$g=1$：実薬群，$g=2$：プラセボ群）

・被験者数：100 例/群

・時点数：ベースライン＋4 時点（時点 4 が主要評価時点）

・測定値（完全データ）の各時点の平均・標準偏差・相関構造：表 11, 12, 図 9

表 11. 完全データの各時点の測定値の平均(SD)

測定値の平均値	ベースライン	時点 1	時点 2	時点 3	時点 4
実薬群	20.0 (4.0)	18.0 (5.0)	15.0 (5.0)	12.0 (6.0)	9.0 (6.0)
プラセボ群	20.0 (4.0)	18.0 (5.0)	16.0 (5.0)	14.0 (6.0)	12.0 (6.0)

表 12. 完全データのベースラインと各時点の測定値の相関

	ベースライン	時点 1	時点 2	時点 3	時点 4
ベースライン	1	0.3	0.3	0.2	0.1
時点 1	—	1	0.6	0.55	0.5
時点 2	—	—	1	0.6	0.55
時点 3	—	—	—	1	0.6
時点 4	—	—	—	—	1

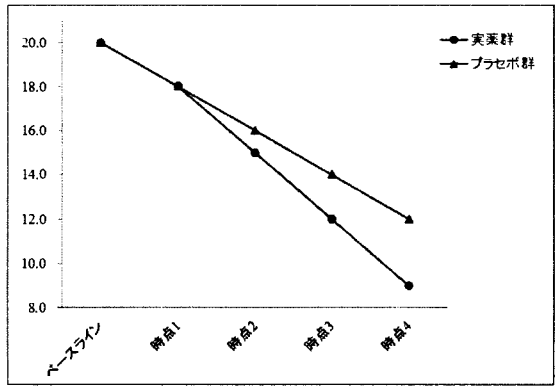


図 9. 各時点の測定値の平均構造（完全データ）

【欠測のあるデータ】

上記の通り作成した完全データを元に， MAR，MNAR の 2 種類の欠測メカニズムを考え，欠測のあるデータを発生させた．なお，時点毎の欠測確率の目標値は以下の値とし，欠測は単調な欠測を仮定した

表 13. 時点毎の欠測確率の目標値

時点 1	時点 2	時点 3	時点 4
5%	10%	13%	15%

上記目標値をもとに作成した欠測確率の関数は下記の通りである．（y_0：ベースライン値，y_i：時点iの測定値，$i=1,2,3,4$）．なお，両群で同じ関数を用いた．

・MAR における時点iの欠測確率　：$p_i = 1 - \dfrac{1}{1+\exp(-5.70+0.14y_{i-1})}$

・MNAR における時点iの欠測確率：$p_i = 1 - \dfrac{1}{1+\exp(-5.70+0.05y_{i-1}+0.10y_i)}$

【検出力・α エラー・MSE 等の算出】

上記の 2 通りの欠測メカニズム（MAR と MNAR）のもとでの欠測のあるデータをそれぞれ 10,000 組ずつ作成し，ベースラインからの変化量に対して MMRM，MI，wGEE，LOCF で解析を行い，検出力，α エラー，MSE（時点 4 における群間差の真値は-3.0）等の算出を行った．なお，有意水準は両側 5%とした．また，α エラー算出の際は，平均構造を両群とも上記プラセボ群の数値に設定した．

7.3 シミュレーションの結果と解釈

シミュレーションの結果を表 14, 15 に示した．まず，α エラーは MAR，MNAR 共に MMRM，MI，LOCF では 5%以下に保たれたが，wGEE は 5%を大きく上回った．検出力は MMRM が最も高く，次いで MI, LOCF, wGEE の順であった．時点 4 の群間差に対する MSE は MMRM が最も小さく，次いで MI，LOCF，wGEE の順であった．最後に時点 4 の群間差の推定値の平均は MMRM, MI, wGEE はほぼ設定値の-3.0 に等しく，LOCF はやや過小評価される傾向がみられた．

以上を踏まえ，本シミュレーションの結果からは MMRM が主解析として望ましいと考えられた．なお，MI もやや劣るものの，ほぼ同様の傾向を示した．また，wGEE は特に α エラーが有意水準の 5%を大きく超えているため，今回のシミュレーションと類似したデータに対する主解析として使用する場合，より詳細な検討が必要であろう．なお，原因としては，重みが極めて大きいデータが全体に大きな影響を与えたことなどが考えられた．LOCF は，α エラーには問題ないものの，検出力が MMRM，MI に比べて劣り，また群間差の推定値を過小評価する可能性が考えられるため，MMRM 等が使用可能ならば，特に使用する必要性は感じられなかった．なお，本解析は 2 シナリオのみのシミュレーションであるため，結果を一般化し過ぎないことに注意が必要である．

表 14. 検出力・α エラー・MSE・点推定値の平均（MAR）

	時点 4 における検出力（%）	時点 4 における α エラー（%）	時点 4 における群間差に対する MSE	時点 4 における群間差の推定値の平均
MMRM	90.40	4.79	0.8250	-2.9988
MI	89.19	4.80	0.8397	-2.9984
wGEE	69.31	11.12	2.3016	-2.9975
LOCF	83.42	4.78	0.9293	-2.8162

表 15. 検出力・α エラー・MSE・点推定値の平均（MNAR）

	時点 4 における検出力（%）	時点 4 における α エラー（%）	時点 4 における群間差に対する MSE	時点 4 における群間差の推定値の平均
MMRM	90.53	4.84	0.8055	-2.9688
MI	89.16	4.70	0.8250	-2.9654
wGEE	69.28	10.40	2.1840	-2.9678
LOCF	86.33	4.76	0.8731	-2.8562

7.4 実際に適用する場合の注意

本シミュレーションでは，MMRM，wGEE，MI の 3 種類の MAR を仮定した解析と LOCF に対して，1 通りのシナリオに対する検討を行った．シミュレーションの結果はデータの分布や欠測メカニズムに大きく影響を受けることが想定されるため，実際に上記と同様のシミュレーションをもとに主解析の選択を行う場合，以下の点などを考慮しつつ，様々なシナリオの検討を行うことが推奨される．

【欠測メカニズム】

① MAR を仮定してよいか？MAR から大きく離れた場合，どの程度影響を受けるか？

【応答変数のデータの分布】

① 外れ値がある場合や，欠測しているデータの分布が観測データと比較的大きく異なる場合どうなるか？

② ベースラインと時点の交互作用等がある場合どうなるか？

【欠測の発生確率】

①欠測が増えるのは，応答変数の値が大きい場合か，小さい場合か？

②補助変数が存在する場合どうなるか？

8. まとめ

本稿では，NRC (2010)の内容を発展させた Mallinckrodt (2013)に従い，"Analytic Road Map"に沿った主解析・感度分析の検討を行った．また，主解析としての MMRM，wGEE，MI，LOCF の性能を 2 通りのシナリオ（欠測メカニズムが MAR と MNAR）のシミュレーションで比較した．

広く言われている通り，欠測の発生メカニズムや解析に与える影響は疾患や薬剤，治験の実施地域等に強く依存する．欠測の生じうる臨床試験の計画・解析の際は，本稿で示したように，シミュレーションによる検討を十分に行って最適な主解析の選択を行い，適切な感度分析を実施した上で，主解析・感度分析の結果を合わせて解釈を行うことが重要である． また，感度分析の研究は現在も極めて盛んに行われている．本稿の内容はあくまで現段階のものであることにも注意されたい．

参考文献

1. 駒嵜弘，高橋文博，横溝孝明．(2014)． 欠測メカニズムに対する感度分析. SAS ユーザー総会論文集.
2. Mallinckrodt, C. H. (2013). *Preventing and Treating Missing Data in Longitudinal Clinical Trials*. Cambridge Press.
3. Mallinckrodt, C. H., Chuang-Stein, C., Molenberghs, G., O'Kelly, M., Ratitch, B., Janssens, M., and Bunouf, P. (2014). Recent development in the prevention and treatment of missing data, *Therapeutic Innovation & regulatory Science* ,**48**, 68-80.
4. Mallinckrodt, C. H., Lane, P. W., Schnell, D., Peng, Y., Mancuso, J. P. (2008). Recommendations for the Primary Analysis of continuous Endpoints in Longitudinal Clinical Trial, *Drug Information Journal*, **42**, 303-319.
5. Mallinckrodt, C. H., Lin, Q., Lipkovich, I., and Molenberghs, G. (2012). A structured approach to choosing estimands and estimators in longitudinal clinical trials, *Pharmaceutical Statistics*, **11**, 456-461.
6. Molenberghs, G., and Kenward, M. G. (2007). *Missing Data in Clinical Studies*. Wiley.
7. National Research Council. (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. The National Academies Press.
8. 大江基貴，土居正明，縄田成毅．(2014)． 解析手法の解説 1. SAS ユーザー総会論文集.
9. O'Kelly, M., and Ratitch, B. (2014). *Clinical Trials with Missing Data -A Guide for practitioners-*. Wiley.
10. 高橋文博，藤原正和，大浦智紀，横山雄一．(2014)．解析手法の解説 2. SAS ユーザー総会論文集.
11. Verbeke, G., and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer.

ODS GRAPHICSを用いた臨床試験データの可視化への挑戦

豊泉 樹一郎, 財前 政美, 北西 由武, 都地 昭夫
塩野義製薬株式会社 解析センター

Challenge to Visualize the Clinical Trial Data with ODS Graphics

Kiichiro Toyoizumi[1)] Masami Zaizen[1)] Yoshitake Kitanishi[1)] Akio Tsuji[1)]
[1)] SHIONOGI & CO., LTD.

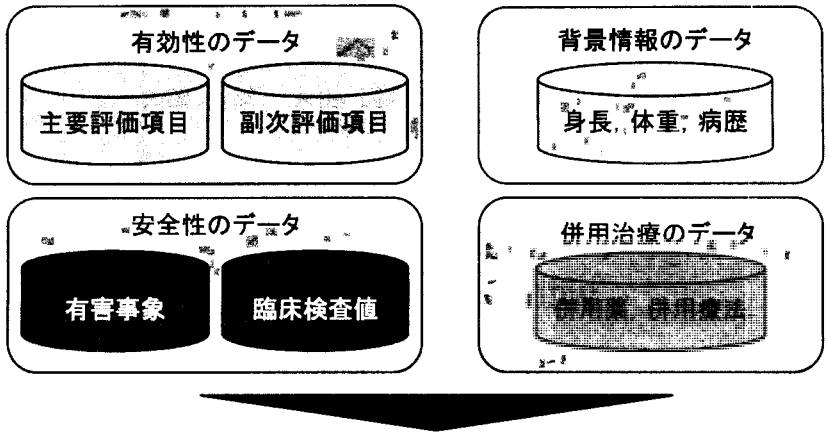
医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

臨床試験の結果は要約統計量や仮説検定によって俯瞰することができる. しかし数値のみの結果を提示するよりも, 適切な可視化を行うことでより理解を深められることが多々ある. 臨床試験における個々の症例プロファイルがその一例である. そこで今回ODS GRAPHICSを用い, 臨床試験データを可視化するためのSASプログラムを開発した. その作成事例について報告する.

キーワード: ODS GRAPHICS

臨床試験のデータ



解析担当者が統計解析を行い，集計結果をまとめる

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

結果の出力例 (主要評価項目の解析)

評価時点	投与群	変化量 調整平均値 (標準誤差)	プラセボとの比較 調整平均値の差 [95%信頼区間]	p 値
2 週	プラセボ	-3.25 (1.27)		
	実薬群	-5.38 (1.35)	-2.13 [-5.78, 1.53]	0.2523
4 週	プラセボ	-6.11 (1.33)		
	実薬群	-8.92 (1.41)	-2.81 [-6.65, 1.02]	0.1499
6 週	プラセボ	-5.51 (1.39)		
	実薬群	-10.86 (1.46)	-5.34 [-9.32, -1.36]	0.0088
10 週	プラセボ	-7.64 (1.46)		
	実薬群	-14.96 (1.53)	-7.32 [-11.51, -3.13]	0.0007

結果の出力例 (有害事象の解析)

器官別大分類 (SOC) 基本語 (PT)	プラセボ N=192 n (%)	件数	実薬群 N=193 n (%)	件数	p 値 [a]
全体	123 (64.1)	210	123 (63.7)	230	1.0000
胃腸障害	43 (22.4)	55	43 (22.3)	64	1.0000
悪心	7 (3.7)	7	14 (7.3)	14	0.1770
便秘	4 (2.1)	4	5 (2.6)	5	1.0000
一般・全身障害および投与部位の状態	9 (4.7)	11	7 (3.6)	19	0.6211
口渇	4 (2.1)	5	5 (2.6)	6	1.0000
倦怠感	4 (2.1)	4	4 (2.1)	4	1.0000

[a] Fisher's exact test.

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

数値のみの結果

- 長所
  - 必要な情報はすべて盛り込まれている
- 短所
  - 一つ一つの結果を見ていくのには時間がかかる

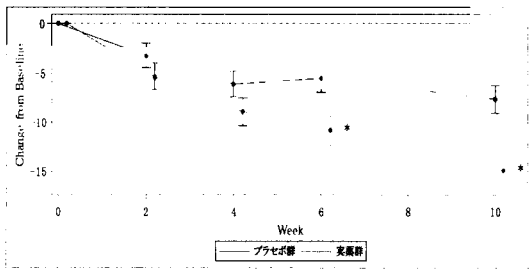
可視化することで，直観的に結果を理解できる
ODS Graphicsで作成

臨床試験の結果で主に見たいもの

- 有効性
  - 主要評価項目の点推定値，区間推定値の経時的な推移
  - どの時点で有意な差があるか
- 安全性
  - プラセボと比較して，どのくらい有害事象のリスクが大きいか（リスク比のプロット）
  - （適用拡大の薬ならば）他の適用症とのリスクの比較
  - 着目している有害事象間の関連性

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

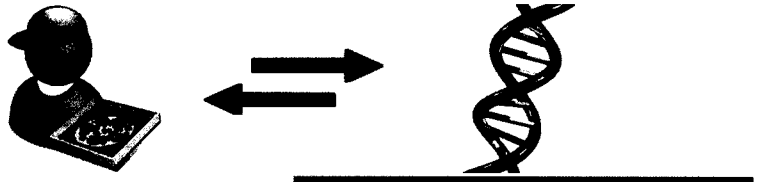
```
proc sgplot data=temp01 noautolegend;
  series x=visit y=chg / name="1" group=TRT;
  scatter x=syx y=syy /markerchar=p markercharattrs=(size=12);
  scatter x=visit y=chg /markerattrs=(symbol=circlefilled)
  yerrorlower=LOW yerrorupper=UPP group=TRT;
  yaxis label="Change from Baseline";
  xaxis label="Week";
  keylegend "1";
  refline 0/axis=y lineattrs=(pattern=2);
run;
```



今までは，グラフに出力される線の種類や色は，Graph Templateで事前に定義しておかなければならなかった。

Graph Template

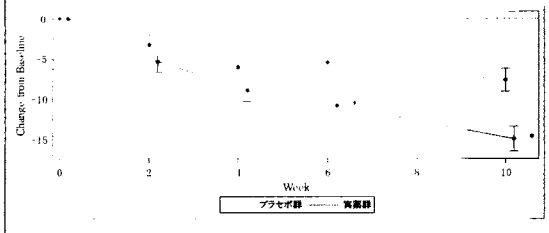
- Graph Template
  - 出力されるグラフの色，線の種類等が定義されている
  - その一つ一つを定義しているのがGraph Template Language (GTL)と呼ばれるSASのコード



ヒト: Graph Template

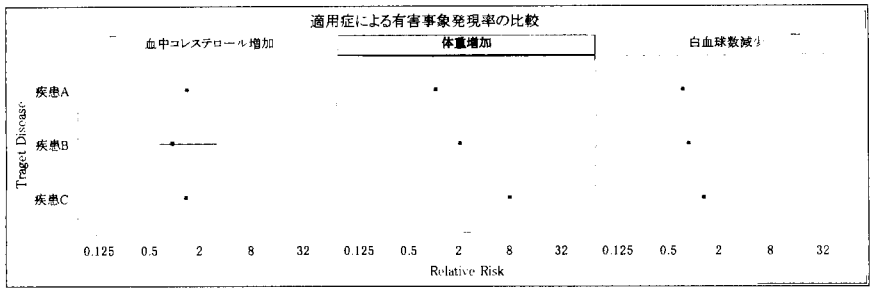
医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

```
proc sgplot data=temp01 noautolegend;
 styleattrs datacontrastcolors=(red green) datalinepatterns=(dot solid);
  series x=visit y=chg / name="1" group=TRT;
  scatter x=syx y=syy /markerchar=p;
  scatter x=visit y=chg /markerattrs=(symbol=circlefilled)
  yerrorlower=LOW yerrorupper=UPP group=TRT;
  xaxis label ="Week";
  yaxis label ="Change from Baseline";
  keylegend "1";
  refline 0/axis=y lineattrs=(pattern=2);
run;
```



SAS 9.4からはstyleattrsステートメントで，線の色，種類などが指定できるようになった

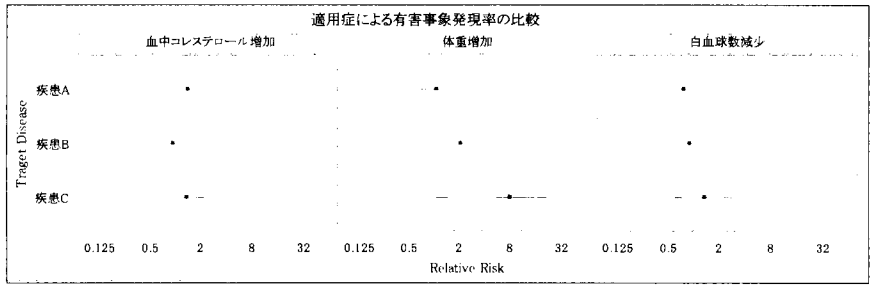
適用症間での有害事象発現のリスク比の比較



各適用症での，プラセボに対する有害事象発現のリスク比とその95%信頼区間をプロットすることで，適用症による有害事象発現率の違いを視覚的に確認できる.

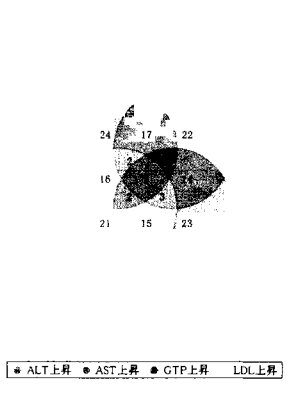
医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

適用症間での有害事象発現のリスク比の比較




```
proc sgpanel data=temp05 noautolegend ;
  panelby  AEDECOD /columns=3 novarname SPARSE LAYOUT= COLUMNLATTICE  ;
  vector x=upp y=disease /xorigin=LOW yorigin=disease group=disease
lineattrs=(pattern=solid) noarrowheads;
  scatter x=estimate y=disease /markerattrs=(symbol=squarefilled);
  COLAXIS LOGSTYLE= LOGEXPAND LOGBASE=2 TYPE= LOG  Label="Relative Risk";
  ROWAXIS values=(1,2,3) min=0.5 max=3.5 valueshint label="Traget Disease";
  refline 1/axis=x lineattrs=(pattern=2);
run;
```

有害事象間の関連(ベン図)




```
ods graphics on/width=10cm height=11cm;
proc sgplot data=temp02 noautolegend;
  styleattrs
  datacolors=(red green blue yellow);
  bubble x=x y=y / bradiusmin=1.7cm
  bradiusmax=1.8cm transparency=0.75
  group=A1 name="1";
  scatter x=x2 y=y2/markerchar=A2;
  xaxis min=-7 max=7 display=none;
  yaxis min=-7 max=7 display=none;
  keylegend "1";
run;
```

どの有害事象と，どの有害事象が重複して発生しているかがわかる

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

ここまでの話

- 試験全体での有効性・安全性の結果の可視化
- 実際には，個々の症例がどのようなプロファイルを辿ったか，一例一例見ていくこともありうる
- 可視化を行わなかった場合，個々の被験者のプロファイルの情報は一覧表から取得することになる

一覧表の例

投与群	患者ID	性別	年齢（歳）	評価時点	実施日	観測値	変化量
プラセボ	001	男	46	ベースライン	2014-03-17	7	0
				2週	2014-03-31	5	-2
				4週	2014-04-14	3	-4
				6週	2014-04-28	3	-4
				10週	2014-05-12	6	-1
				13週	2014-06-09	3	-4
				14週	2014-06-09	3	-4
薬剤群	002	女	57	ベースライン	2014-04-01	8	0
				2週	2014-04-15	4	-4

投与群	患者ID	性別/年齢（歳）	基本語（報告名）	発現時期 [a]/持続期間 [b]	重症度/重篤区分/因果関係	治験薬の処置/治験薬以外の処置/転帰	コメント
プラセボ	001	男 46	口内炎（口内炎）	5 16	軽度 非重篤 関連なし	投与量変更せず いいえ 回復	
			血中コレステロール増加（血中コレステロール増加）	15 96	中等度 非重篤 関連なし	投与量変更せず いいえ 回復	
			便秘（便秘）	35 56	軽度 非重篤 関連なし	投与量変更せず いいえ 回復	
			回転性めまい（回転性めまい）	65 4	重度 非重篤 関連あり	投与量変更せず いいえ 回復	

[a] 発現時期（日）=（発現日）-（初回服薬日）+1
[b] 持続期間（日）=（転帰日）-（発現日）+1

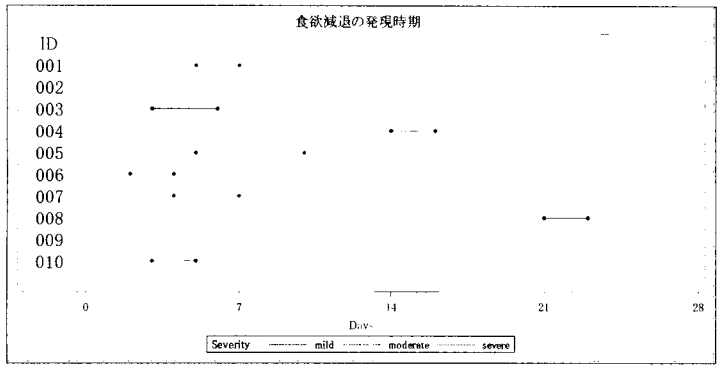
各症例について1個1個見ていくのは大変
→可視化が重要

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

個別症例のプロットで見たいもの

- 安全性
  - 着目したい有害事象が治験薬投与後のどの時点から発現しているのか＆どのくらいの期間継続して発現しているのか
- 有効性
  - 各被験者での死亡や症状悪化といったイベントがどの時点で発生しているか
  - 主要評価項目の経時的な推移に対する有害事象や併用薬の及ぼす影響

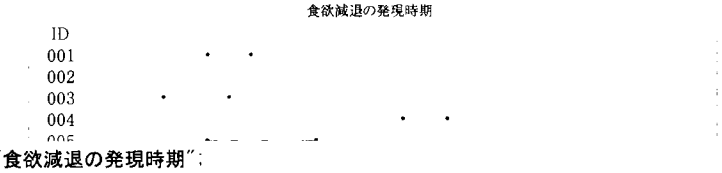
特定の有害事象の各症例での発現時期



・各被験者で，どの時点で食欲減退が発生しているのかが分かる
・プロットの色を変えることで，有害事象の重症度の違いを表すことも可能

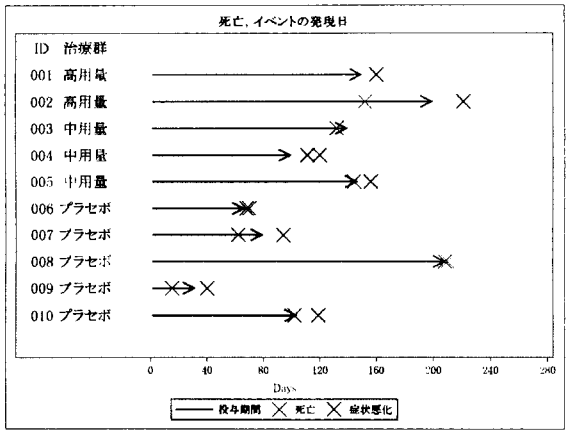
医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

特定の有害事象の各症例での発現時期




```
title "食欲減退の発現時期";
proc sgplot data=temp03 noautolegend;
  styleattrs  datacontrastcolors=(blue green red);
  vector x=AEENDY y=seq/ xorigin=AESTDY yorigin=seq group=AESEVC
  lineattrs=(pattern=solid thickness=2) noarrowheads name="1" nomissinggroup;
  scatter x=AESTDY y=seq / markerattrs=(symbol=circlefilled color=black size=5pt) ;
  scatter x=AEENDY y=seq / markerattrs=(symbol=circlefilled  color=black  size=5pt) ;
  scatter x=dummy y=seq  /markerchar=subjid markercharattrs=(size=14) x2axis;
  xaxis label="Days" offsetmin=0.10 offsetmax=0.01 values=(0 to 28 by 7);
  x2axis offsetmin=0.01 offsetmax=0.92 min=0.5 max=1.5 values=(1) display= NONE;
  yaxis values=(-11 to 0 by 1) label="被験者番号" display= NONE;
  keylegend "1" "2"/title="Severity";
run;
```

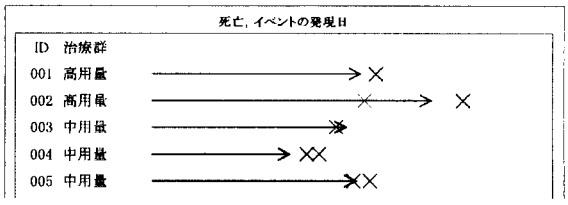
各症例でのイベント発現日一覧



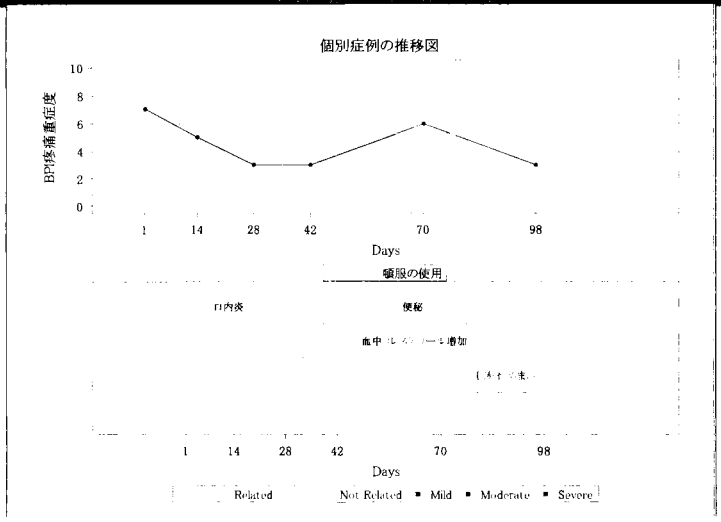
・どの症例で，いつ死亡や症状悪化が起こったか，治験薬がどのくらいの期間投与されていたかが視覚的にわかる

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

各症例でのイベント発現日一覧

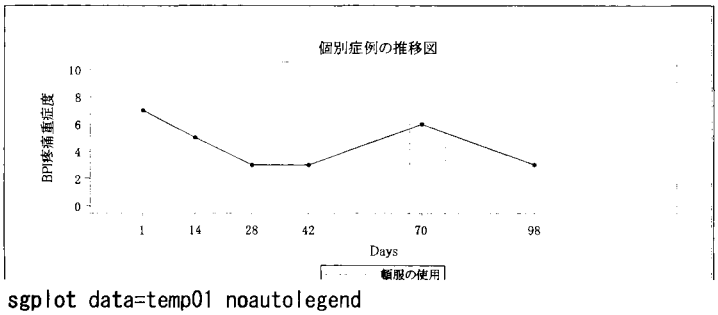



```
title "死亡，イベントの発現日";
proc sgplot data=temp04 noautolegend;
  vector x=EXENDY y=seq/ xorigin=EXSTDY yorigin=seq  lineattrs=(pattern=solid thickness=2
color=balck)  name="1" legendlabel="投与期間";
  scatter x=DEATH y=seq / markerattrs=(symbol=X color=red size=14pt) name="2" legendlabel="死亡" ;
  scatter x=RISK y=seq  / markerattrs=(symbol=X color=blue  size=14pt) name= "3"  legendlabel= "症状
悪化";
  scatter x=x1 y=seq/markerchar=subjid markercharattrs=(size=14) x2axis;
  scatter x=x2 y=seq/markerchar=TRTPC markercharattrs=(size=14) x2axis;
  xaxis label="Days" offsetmin=0.25 offsetmax=0.01 values=(0 to 280 by 40);
  x2axis offsetmin=0.01 offsetmax=0.75 min=0.5 max=3.5 values=(1,2) valueshint display= NONE;
  yaxis values=(-11 to 0 by 1) label="被験者番号" DISPLAY= NONE;
  keylegend "1" "2" "3";
run;
```

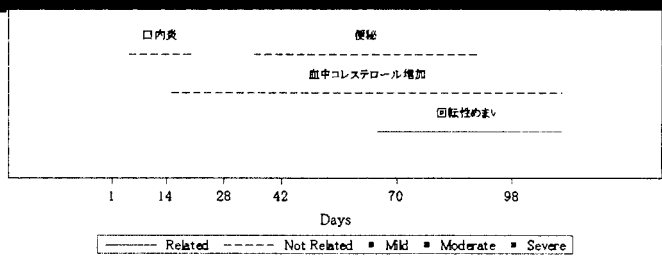


・各被験者での主要評価項目の推移やどの時点でレスキュー投与されたかがわかる
・有害事象の重症度や因果関係を，線種や色で表すことも可能

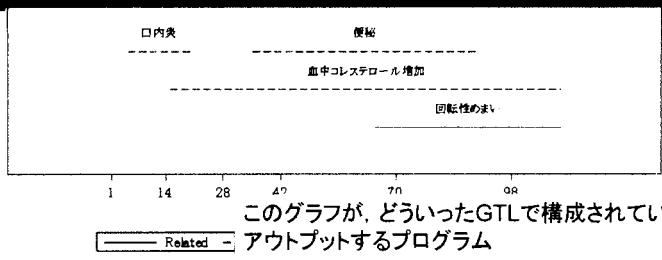
医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014




```
proc sgplot data=temp01 noautolegend
  tmplout="C:¥Users¥s99999¥Desktop¥ex1.sas";
  series x=ADY y=AVAL;
  scatter x=ADY y=AVAL/markerattrs=(symbol=circlefilled);
  xaxis values=(1,14,28,42,70,98) max=120 valueshint label="Days";
  yaxis values=(0 to 10 by 2)  label="BPI疼痛重症度";
  refline CMSTDY/axis=x legendlabel= "頓服の使用"
  lineattrs=(pattern=2 color=red) name="1";
  keylegend "1";
run;
```




```
proc sgplot data=temp01 noautolegend tmplout="C:¥Users¥s99999¥Desktop¥ex2.sas";
  vector x=AEENDY1 y=seq/xorigin=AESTDY1 yorigin=seq noarrowheads name="2"
  legendlabel="Related" lineattrs=(color=black pattern=solid);
  vector x=AEENDY2 y=seq/xorigin=AESTDY2 yorigin=seq noarrowheads name="3 "
  legendlabel="Not Related" lineattrs=(color=black pattern=mediumdash) ;
  scatter x=AEMED y=seq2/markerchar=AEDECOD1 markercharattrs=(color=black) name="4 "
  legendlabel="Mild";
  scatter x=AEMED y=seq2/markerchar=AEDECOD2 markercharattrs=(color=blue) name="5 "
  legendlabel="Moderate" ;
  scatter x=AEMED y=seq2/markerchar=AEDECOD3 markercharattrs=(color=red)  name="6 "
  legendlabel="Severe";
  xaxis values=(1,14,28,42,70,98) min=-10 max=120 label="Days" valueshint;
  yaxis values=(0 to 4 by 1) display=none ;
  keylegend "2" "3" "4" "5" "6"  / position=bottom;
run;
```

このグラフが，どういったGTLで構成されているのかを
アウトプットするプログラム

```
proc sgplot data=temp01 noautolegend tmplout="C:¥Users¥s99999¥Desktop¥ex2.sas";
  vector x=AEENDY1 y=seq/xorigin=AESTDY1 yorigin=seq noarrowheads name="2"
  legendlabel="Related" lineattrs=(color=black pattern=solid);
  vector x=AEENDY2 y=seq/xorigin=AESTDY2 yorigin=seq noarrowheads name="3 "
  legendlabel="Not Related" lineattrs=(color=black pattern=mediumdash) ;
  scatter x=AEMED y=seq2/markerchar=AEDECOD1 markercharattrs=(color=black) name="4 "
  legendlabel="Mild";
  scatter x=AEMED y=seq2/markerchar=AEDECOD2 markercharattrs=(color=blue) name="5 "
  legendlabel="Moderate" ;
  scatter x=AEMED y=seq2/markerchar=AEDECOD3 markercharattrs=(color=red)  name="6 "
  legendlabel="Severe";
  xaxis values=(1,14,28,42,70,98) min=-10 max=120 label="Days" valueshint;
  yaxis values=(0 to 4 by 1) display=none ;
  keylegend "2" "3" "4" "5" "6"  / POSITION=bottom;
run;
```

ex1.sasの中身を見てみると・・・

```
proc template;
define statgraph sgplot;                    ①有効性の推移図を作成しているGTL
begingraph /;
layout overlay / xaxisopts=( Label="Days" type=linear
linearopts=( tickvaluelist=( 1 14 28 42 70 98 ) viewmax=120 ) )
y2axisopts=(labelFitPolicy=Split) yaxisopts=( Label="BPI疼痛重症度"
labelFitPolicy=Split type=linear linearopts=( tickvaluelist=( 0 2 4 6 8
10 ) viewmin=0 viewmax=10 ) ) y2axisopts=(labelFitPolicy=Split);
   seriesPlot X=ADY Y=aval / legendlabel="aval" name="SERIES";
   scatterPlot X=ADY Y=aval / primary=true
markerattrs=( Symbol=CIRCLEFILLED) legendLabel="aval" name="SCATTER";
   ReferenceLine X=CMSTDY / clip=true name="1" legendLabel="頓服の使用"
lineattrs=( Color=CXFF0000 Pattern=2);
   discreteLegend "1" / location=Outside;
endlayout;
endgraph;
end;
run;
```

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

有効性の推移図と有害事象の推移図を合成したGTLの作成

```
proc template ;
  define statgraph surface;
  dynamic _ticklist_ ;          どのくらいの比率で、①と②を配合するのかを定義
  begingraph ;
  entrytitle "個別症例の推移図" /;
  layout lattice /
  backgroundcolor=white rows=2 rowweights=(.5 .5) order=columnmajor
  pad=20px border=no ;
```

①有効性の推移図を作成しているGTL(前頁の赤枠の部分)を貼り付ける
→次頁の赤枠部分

② ①と同様に有害事象の推移図を作成しているGTLを貼り付ける

```
endlayout;
endgraph;

  end;
run;
```

```
proc template ;
 define statgraph surface;
  dynamic _ticklist_ ;
  begingraph ;
   entrytitle  "個別症例の推移図" /;
   layout lattice /
   backgroundcolor=white rows=2 rowweights=(.5 .5) order=columnmajor pad=20px border=no
layout overlay / xaxisopts=( Label="Days" type=linear linearopts=( tickvaluelist=( 1 14 28 42 70 98 ) viewmax=120 ) )
y2axisopts=(labelFitPolicy=Split) yaxisopts=( Label="BPI疼痛重症度" labelFitPolicy=Split type=linear
linearopts=( tickvaluelist=( 0 2 4 6 8 10 ) viewmin=0 viewmax=10 ) ) y2axisopts=(labelFitPolicy=Split);
   seriesPlot X=ADY Y=aval / legendLabel="aval" name="SERIES";
   scatterPlot X=ADY Y=aval / primary=true markerattrs=( Symbol=CIRCLEFILLED) legendLabel="aval" name="SCATTER";
   referenceLine X=CMSTDY / clip=true name="1" legendLabel="頓服の使用" lineattrs=( Color=CXFF0000 Pattern=2);
   DiscreteLegend "1" / Location=Outside;
endlayout;
layout overlay / x2axisopts=(labelFitPolicy=Split) xaxisopts=( Label="Days" labelFitPolicy=Split type=linear
linearopts=( tickvaluelist=( 1 14 28 42 70 98 ) viewmin=-10 viewmax=120 ) ) yaxisopts=( display=none type=linear
linearopts=( tickvaluelist=( 0 1 2 3 4 ) viewmin=0 viewmax=4 ) ) x2axisopts=(labelFitPolicy=Split);
   VectorPlot X=AEENDY1 Y=seq XOrigin=AESTDY1 YOrigin=seq / Lineattrs=( Color=CX000000 Pattern=1) Arrowheads=false
LegendLabel="Related" NAME="2";
   VectorPlot X=AEENDY2 Y=seq XOrigin=AESTDY2 YOrigin=seq / Lineattrs=( Color=CX000000 Pattern=4) Arrowheads=false
LegendLabel="Not Related" NAME="3";
   ScatterPlot X=AEMED Y=seq2 / primary=true MarkerCharacter=AEDECOD1 MarkerCharacterAttrs=( Color=CX000000)
LegendLabel="Mild" NAME="4";
   ScatterPlot X=AEMED Y=seq2 / MarkerCharacter=AEDECOD2 MarkerCharacterAttrs=( Color=CX0000FF) LegendLabel="Moderate"
NAME="5";
   ScatterPlot X=AEMED Y=seq2 / MarkerCharacter=AEDECOD3 MarkerCharacterAttrs=( Color=CXFF0000) LegendLabel="Severe"
NAME="6";
   DiscreteLegend "2" "3" "4" "5" "6" / Location=Outside valign=bottom;
endlayout;
endlayout;
endgraph;
  end;
run;
```

どのくらいの比率で，①と②を配合するのかを定義

①有効性の推移図を構成しているGTL

②有害事象の推移図を構成しているGTL

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

SGRENDERを用いて出力

```
ods rtf file="C:¥Users¥s99999¥Desktop¥patient_profile.rtf" bodytitle;
%macro patient_profile(subjid);
title "SUbject ID: &subjid.";
proc sgrender data=temp01 template=surface;
where subjid="&subjid.";
run ;
%mend patient_profile;

data _null_;
  set temp02;
  call execute('%patient_profile('||strip(subjid)||');');
run;
ods rtf close;
```

作成したTemplateをもとに，SGRENDER，Call Executeで
出力することで個々人の被験者ごとの，推移図が作成される

拡張

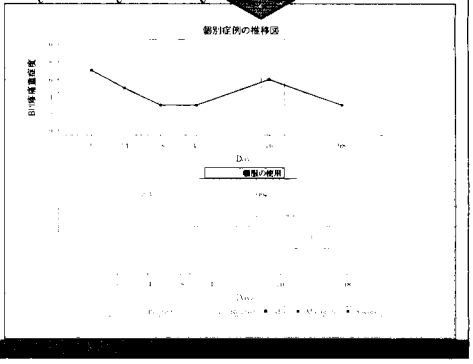
- これらの個別症例のプロファイルや，背景情報などを結合したファイルの作成
- ハイパーリンクや見出しをつけて作成することで，簡単に各被験者がどのような推移を辿ったのか情報を見ることが可能になる

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

Example

Demo Data

ID	投与群	完了例	年齢	性別	併用療法	併用薬	AE	Serious AE	ADR
001	実薬群	Y	62	Male	Y	Y	Y		
003	Placebo	Y	53	Male	Y	Y	Y		
004	Placebo	N	48	Male	Y	Y	Y		
006	実薬群	Y	49	Female	Y	Y	Y		
007	実薬群	Y	48	Male					



個別症例のファイルを作成するメリット

- 探索解析を行う場合，有効性・安全性について個々の症例のプロファイルの検討から始めることが少なくない
- 実際の個々の症例のデータを見ながら，探索解析のアイディアを思いつくこともある
- しかし数値データのみから，アイディアを思いつくのは困難

可視化が重要！

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

Conclusion

- 臨床試験のデータにおける可視化の重要性
  - 臨床試験の結果について，数値で集計することは必要かつ重要
  - しかし，数値結果だけでは直観的な考察を得ることは困難
- SG ProcedureとODS Graphicsを用いることで，個々の症例のプロファイルなど様々な視点からの可視化が可能になる
- ODS Graphicsによる可視化により，新たな知見，アイディアを得ることが期待される

End Of Slide

投与前値を含むクロスオーバー法での経時データの解析

高橋 行雄
BioStat 研究所(株)

Analysis of Longitudinal Data on Crossover Designs with Baseline

Yukio Takahashi
BioStat Research Co.,Ltd.

要旨： 投与前値を含む経時データの群間比較において，投与前値を共変量とする線形混合モデルによる解析，あるいは，投与前値を各群で共通とした解析が，平均への回帰現象を受けにくく，投与前値からの差による群間比較より統計的に優れていることが知られている．クロスオーバー法による臨床薬理試験では，投与前値を含む経時データがしばしば得られる．同一症例内の各実験時期の投与前値の変動が少ないとみなせるのであれば，投与後のデータのみ，あるいは，前後差での解析を行うことの妥当性がある．しかしながら，同一症例内で投与前値に無視できない変動があり，また投与後値との間でなんらかの相関関係があった場合には，投与前値を共変量とした解析を行う必要がある．各種の統計モデルを適用した際に，薬剤群間の差の信頼区間がどのように変化するのかを，SAS for Mixed Model 2ed.（Little（2006））5.1 節で例示されているクロスオーバー法による 1 秒あたりの呼吸量の経時データを用いて比較する．さらに，それぞれの統計モデルによる誤差構造についての比較も行う．

キーワード： QT/QTc 試験，経時データ，MIXED プロシジャ，共分散分析，クロスオーバー法

1. はじめに

ICH-S7b「ヒト用医薬品の心室再分極遅延（QT 間隔延長）の潜在的可能性に関する非臨床的評価」が，2009 年 10 月 23 日にステップ 5 となり，大動物を使ったクロスオーバー法によるテレメトリーQT/QTc 試験が多くの研究施設で行われるようになってきた．この試験から複数の投与前値を含む超多時点の経時データが得られる．このようなデータに関して，どのような経時データの解析を行うかについては適当な文献・成書がない．

ICH-S7b には，統計解析についての具体的な記載がないので，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（2009）」を参考にする．ICH-E14 の 2.2 節 QT/QTc 評価試験に「QT/QTc 評価試験の目的は、被験薬に心室再分極に対する一定の大きさ以上の薬理作用があるか否かを決定することであり、その値は QT/QTc 間隔の延長として検出される。規制当局が関心をもつ基準値レベルについては後述するが、QTc 間隔への作用の平均値としておよそ 5ms であり、95%信頼区間の上限を 10ms とするものである。」と信頼区間方式による判定基準が示されている．

2.2.4節の QT/QTc評価試験の解釈には，「同様の考え方に基づき、QT/QTc評価試験が陰性とは、その薬剤の QTc 間隔への時間を一致させた平均効果の最大値に対する 95%片側信頼区間の上限が10ms を下回る場合を指す。この定義は、被験薬の QT/QTc 間隔への作用の平均がおよそ 5ms を超えないことを合理的に保証するために選択されている。時間を一致させた差の最大値がこの基準値を超える場合、試験結果は陽性とされる。」と判定基準を定めている．

ベースライン値については，3.2 節 QT/QTc 間隔データの解析に「QT/QTc 間隔のベースラインに比しての延長は注意すべき徴候であるが、それらは平均値への回帰や極端な値を選択したためなど薬物療法に無関係な要因による変化である可能性があるので、QT/QTc 間隔のベースラインとの差の解釈は複雑である。」と注意している．さらに，「QT/QTc 間隔データは、中心傾向（central tendency）の解析（例えば、平均値、中央値）及びカテゴリカル解析の両方の形で示すべきである。どちらの解析も、臨床上のリスクを評価する際の適切な情報となり得る。」と中心傾向の解析の必要性が強調されている．

中心傾向とは，質問紙を用いた評価では「どちらともいえない」といったようなほぼ中心に回答が集まる現象として知られている． なお，3.2.1 節で「被験薬が QT/QTc 間隔へ与える作用の解析は、最も一般的には、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。」と述べられ，いくつかの例示もあるが，漠然としていてどのような解析なのかが不明瞭である．「平均値への回帰」については，丁寧な説明が別にされているので，一般的に用いられている投与前値を共変量とした解析でもないようである．

ICH-E14 の Q&A 問 6 に「ベースライン値の必要性について説明して下さい。また、ベースライン値が必要な場合、QT/QTc 評価試験がクロスオーバー試験と並行群間比較試験のデザインで実施されるそれぞれの場合について、ベースライン値の測定方法を説明して下さい。」とあり，回答では，被験薬の投与に先立って同時刻に測定されたベースライン，投与前のベースラインの 2 つがあり，スロスオーバー法では投与前のベースライを用いることが適切と述べている．

これらの文脈から，「ベースライン値による調整」とは，クロスオーバー法では，投与前値からの差であることが推測され，投与前値を共変量とした調整ではないようである．

そこで，Littell ら(2006)が SAS for Mixed Model 2ed. 5.1 節で例示しているクロスオーバー法による 1 秒あたりの呼吸量の経時データを用いて，各種の統計モデルを適用した際の，主要評価時点における薬剤群間の差の信頼区間について比較検討し，テレメトリーQT/QTc 試験の経時データの解析法の参考とすることにした．

2. データの構造と判定基準

2.1 データの概要

Littell らの呼吸機能の経時データを 表 1 に示す．これは，呼吸機能の改善を目的にした薬剤 T について，標準薬 S とプラセボ P を対象にした 24 症例のクロスオーバー試験の結果である．各群について，1 秒あたりの呼吸量 FEV1（Forced Expiratory Volume in 1st second）が単回投与前から 8 時間後まで 1 時間ごとに測定されている．なお，文献では 24 症例に対して 3 種の薬剤をランダムに割り付けたと述べられているだけで，実験順序・時期などのデータは含まれていない．

表 1　FEV1 の経時変化

drug	patient	0	1	2	3	4	5	6	7	8
T(a)	201	2.46	2.68	2.76	2.50	2.30	2.14	2.40	2.33	2.20
T(a)	202	3.50	3.95	3.65	2.93	2.53	3.04	3.37	3.14	2.62
	:									
T(a)	232	2.49	3.73	3.51	3.16	3.26	3.07	2.77	2.92	3.00
S(c)	201	2.30	3.41	3.48	3.41	3.49	3.33	3.20	3.07	3.15
S(c)	202	2.91	3.92	4.02	4.04	3.64	3.29	3.10	2.70	2.69
	:									
S(c)	232	2.79	4.10	3.85	4.27	4.01	3.78	3.14	3.94	3.69
P(p)	201	2.14	2.36	2.36	2.28	2.35	2.31	2.62	2.12	2.42
P(p)	202	3.37	3.03	3.02	3.19	2.98	3.01	2.75	2.70	2.84
	:									
P(p)	232	2.88	3.04	3.00	3.24	3.37	2.69	2.89	2.89	2.76

http://support.sas.com/publishing/bbu/59882/59882.zip からダウンロードし整形.

表 1 に示した 3 例についての 図 1 に示す線グラフで経時変化の特徴を概観する．症例により投与前値が異なり，同じ症例の中でも各薬剤の投与前値に症例間ほどではないが差があり，プラセボ投与以外は最初の 1 時間目から反応があり，8 時間目まで継続していることが観察される．

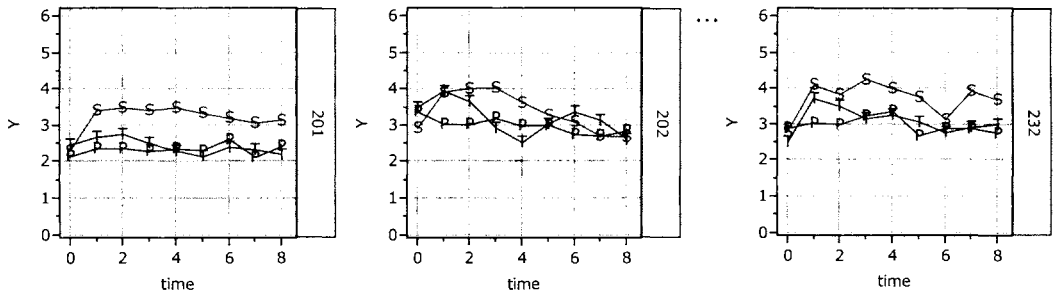


図 1　症例ごとの FEV1 の経時変化

全症例の経時変化の平均値 (y) の推移および投与前からの差 (d) について 図 2 に示す．S 薬および T 薬の投与後 1 時間目で反応がピークとなり，その後 8 時間目まで緩やかに減少している．P 群では大きな変動は見られない．

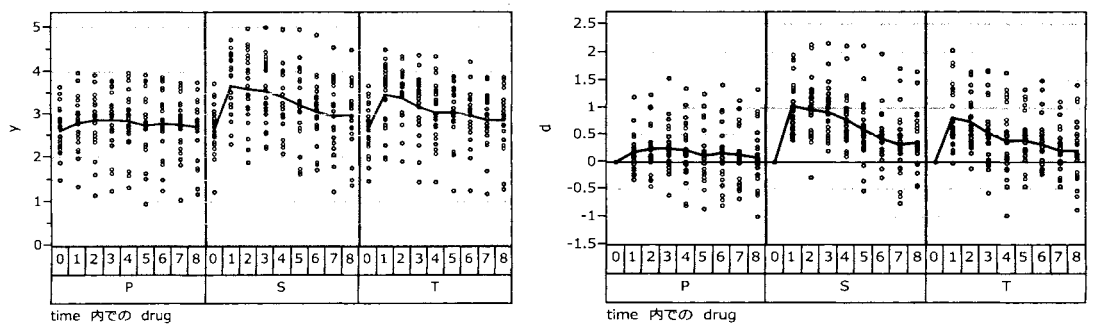


図 2　症例別 FEV1 の群ごとの経時変化

2.2 信頼区間方式による薬効の判定基準

この試験は薬効評価のための試験であり，有効性判定基準がないため，QT/QTc 評価試験に準じた信頼区間方式による判定基準を次のように別途定める．

主要な評価時期に対して，試験薬 T の片側 95%信頼区間の下限が，プラセボ P の同時期の点推定に対して 15%増の反応であれば，薬効がある用量とみなす．また，陽性対照である標準薬 S に対しても同様の基準を適用し，これを満たした場合はこの試験が適切に行われたたと判断する．なお，試験薬 T を複数用量設定すれば，標準薬 S の投与量に対する等価用量の推定が可能となる．

並行群間試験の場合には，投与前値からの評価時点までの差 d を主要変数とする場合には，投与前値を共変量として解析モデルに含めること．これは，主要評価時点の反応を y とし投与前値を共変量とした場合の薬剤間に関する解析結果と一致することが知られているためである．

なお，クロスオーバー法による実験の場合には，各症例のそれぞれの薬剤群の投与前値を共変量としなくとも薬剤群間差の推定が適切に行える可能性もあるので，投与前値を共変量として含める場合には，含めない場合の結果を示し，総合的な判断をすることが必要である．

また，測定時点を含めた解析を追加する場合には，主要評価時点での薬剤群間差に関する各種の推定を行い，総合的な薬効の判定に加えても差し支えない．

3. 主要変数についての解析

3.1 主要変数についての解析

クロスオーバー法では，症例を変量効果，薬剤を固定効果とした解析が適切とも思われるが，生物学的同等(BE)試験では，症例を固定効果とした解析が定式化されている．そこで，両者での解析結果を比較し，結果の判定にどのような影響をあたえるのかを検討した．

反応のピークとなる 1 時間目のデータを対象として，症例，および薬剤を固定効果とした 2 元配置繰り返しなしの分散分析表を 表 2 示す．「平均平方の構造」は，GLM プロシジャの RANDOM ステートメントで症例を指定することによって得られる．この構造から症例についての分散成分を計算した結果を付け加えてある．

症例を変量効果とした MIXED プロシジャの REML 法（制限付き最尤法）による解析では，変量効果に関する分散成分の出力と固定効果とした薬剤の自由度および F 値 のみが出力されるが伝統的な 表 2 の様式の分散分析表の出力はない．

表 2　1 時間目における 2 元配置とした分散分析表と分散成分

要因	自由度	平方和	平均平方	F 値	p 値	平均平方の構造	分散成分
patient	23	27.1613	1.1809	13.02	<.0001	$\sigma_e^2+3\sigma_{patient}^2$	0.3634
drug	2	9.9948	4.9974	55.08	<.0001	$\sigma_e^2+24\sigma_{drug}^2$	
誤差	46	4.1734	0.0907			σ_e^2	0.0907
全体	71	41.33					

症例を固定効果とするか，変量効果とするかで，表 3 に示すように各薬剤の信頼区間に異なる結果を与える．プラセボ群の平均値は 2.8150 の 15% 増は 3.2373 であり，であり，症例を固定効果とした場合の S 薬の信頼区間の 90%下限は 3.5818，T 薬は 3.3855 と大きく離れている．他方，症例を変量効果とみなした場合に T 薬の信頼区間の 90%下限は 3.2553 と下方に広がり，T 薬の場合には

プラセボ群の平均値の 15%増である 3.2488 をわずかに上回る結果である.

表 3　1 時間目における 2 種類の *SE* および信頼区間

			症例を固定効果		症例を変量効果			
薬剤		平均	*SE*	*L* 90%	*SE*	*L* 90%		P群の15%増
P		2.8150	0.0615	2.7118	0.1376	2.5816		
S		3.6850	0.0615	3.5818	0.1376	3.4516		3.2373
T		3.4888	0.0615	3.3856	0.1376	3.2553		3.2373
薬剤	薬剤	差	*SE*	*L* 90%	*SE*	*L* 90%	*p* 値	P群の15%
S	P	0.8700	0.0870	0.7240	0.0870	0.7240	<.0001	0.4223
T	P	0.6738	0.0870	0.5278	0.0870	0.5278	<.0001	0.4223
S	T	0.1963	0.0870	0.0503	0.0870	0.0503	0.0288	

症例を固定効果した場合，変量効果とした場合の薬剤群の信頼区間の 90%下限が異なるのは，以下に示すように症例に関する分散成分を加味するかしないかによって説明される.

症例を変量効果とした場合の分散成分は，平均平方の構造から症例の分散 $\hat{\sigma}^2_{patient}$ は,

$$\hat{\sigma}^2_{patient} = (V_{patient} - V_e)/3 = (1.1809 - 0.0907)/3 = 0.3634$$

と推定できる．症例を固定効果とした場合に，各薬剤の固定 SE_{drug} は，24 症例の平均値に対するものなので,

$$\text{固定}\, SE_{drug} = \sqrt{V_e/24} = \sqrt{0.0907/24} = 0.0615$$

である．これに対し，症例を変量とした場合には,

$$\text{変量}\, SE_{drug} = \sqrt{(V_e + \hat{\sigma}^2_{patient})/24} = \sqrt{(0.0907 + 0.3634)/24} = 0.1376$$

誤差分散（誤差の平均平方と同じ）に症例に関する分散成分を加えた結果となる．症例を固定効果とした場合の *SE* は，同一の症例を対象として実験を繰り返した場合の各薬剤の母平均値に関するものであり，症例を変量効果とした場合の *SE* は，別の症例を対象にした場合となっていて，一般化可能性の観点からは，症例を変量効果とした場合の *SE* から信頼区間を計算することが望ましい.

薬剤間の差 SE_{drug} は，症例を固定効果とした場合でも，変量効果とした場合でも,

$$\text{群間差}\, SE_{drug} = \sqrt{2V_e/24} = \sqrt{2 \times 0.0907/24} = 0.0870$$

と同じである．これは，異なる症例に対する実験であっても，各々の症例の反応の大きさに違いがあったとしても，同じ症例内での薬剤間の比較なので，症例に関する分散成分が入り込まないからである．プラセボ群の平均値の 15%は 0.4223 であり，T 薬の差の信頼区間の 90%下限は 0.5278 であるので，ゆとりをもって T 薬の薬効が証明されたことになる.

プラセボの平均値の 15%増増による評価と，プラセボと T 薬の差がプラセボの平均値の 15%増を用いた判定と 2 通りが考えられる．どちらが適切なのだろうか.

3.2　投与前値からの差での解析

投与前値からの差（変化量）による解析は，元データに比べて変化の大きさが明確で，相対的な薬効の比較がしやすいとの利点もあり，また投与前値の症例間の変動を除去できるために元データ

での解析よりも望ましいのではないかと思われている．クロスオーバー法の場合は，固定効果，あるいは変量効果として症例をモデルに組み込んでいるので， *SE* がどのように変化するのか検討する．

図 3 に示すように投与前と投与後の 1時間目のFEV1の間の相関は，プラセボ群で相関係数0.82，S 薬 0.74，T 薬 0.61，といずれも 0.5 以上であるので，投与前からの差での解析が，元データでの解析よりも望ましいとも思われる．これは，並行群間試験の場合であり，クロスオーバー試験でも成り立つかを検討する．

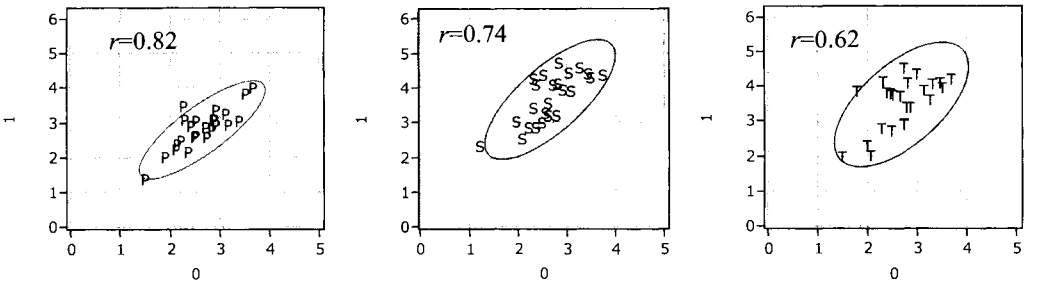


図 3　薬剤群ごとの投与前値と 1 時間目の相関関係

表 4 にクロスオーバー法ではなく並行群間試験とみなした結果を示す．元データでの T 薬の平均は 3.4888，*SE* は 0.1376 であるのに対し，投与前からの差のデータでは，T 薬の差の平均は 0.8204，*SE* は 0.0973 と元データに比べて *SE* が小さくなっている．プラセボと T 薬との差の *SE* は，元データで 0.1945 であり，信頼区間の 90%下限は 0.3494 と P 群の 15%よりも小さいので，薬効があるかは判定保留となる．差の差のデータの *SE* は 0.1376 で，信頼区間の 90%下限は 0.4119 と P 群の 15%よりも小さいので，薬効があるかは判定保留となる．

並行群間試験とみなした解析は，症例間の変動が信頼区間の 90%下限の計算に含まれるために，厳しい判定となってしまうが，投与前からの差をとることによって *SE* が小さくなり，信頼区間の 90%下限が狭まることが確認された．

表 4　並行群間試験とみなした解析結果

		元データでの群間比較			差のデータでの群間比較			
薬剤		平均	*SE*	*L* 90%	差の平均	*SE*		P群の15%増
P		2.8150	0.1376	2.5857	0.1792	0.0973		
S		3.6850	0.1376	3.4557	1.0413	0.0973		3.2373
T		3.4888	0.1376	3.2594	0.8204	0.0973		3.2373
薬剤	薬剤	差	*SE*	*L* 90%	差の差	*SE*	*L* 90%	P群の15%
S	P	0.8700	0.1945	0.5457	0.8621	0.1376	0.6327	0.4223
T	P	0.6738	0.1945	0.3494	0.6413	0.1376	0.4119	0.4223
S	T	0.1963	0.1945	-0.1281	0.2208	0.1376	-0.0086	

投与前からの差のデータに対してクロスオーバー法での解析結果を 表 5 に示す．症例を固定効果とした場合には，各薬剤の *SE* は 0.0642 と
表 4 差の *SE* 0.0973 よりかなり小さい．しかしながら，症例を変量効果とした場合の *SE* は 0.0973 と同程度であるが，プラセボと T 薬の平均値の差の *SE* は固定効果でも変量効果でも同じ 0.0883 で

あり並行群間試験の場合の 0.1376 にくらべて明らかに小さくなっている.

表 5　投与前値からの差についての信頼区間の 90%下限値

			症例を固定効果		症例を変量効果			
薬剤		差の平均	*SE*	*L* 90%	*SE*	*L* 90%		
P		0.1792	0.0624	0.0744	0.0973	0.0154		
S		1.0413	0.0624	0.9365	0.0973	0.8775		
T		0.8204	0.0624	0.7156	0.0973	0.6567		
薬剤	薬剤	差の差	*SE*	*L* 90%	*SE*	*L* 90%	*p* 値	P群の15%
S	P	0.8621	0.0883	0.7139	0.0883	0.7139	<.0001	0.4223
T	P	0.6413	0.0883	0.4930	0.0883	0.4930	<.0001	0.4223
S	T	0.2208	0.0883	0.0726	0.0883	0.0726	0.016	

投与前からの差での解析において，クロスオーバー法の良さは実感できたのであるが，表 3 に示した元データでの症例を変量効果とした場合の解析結果と比較してみよう．元データでの T 薬の *SE* は 0.1376 であるが，差のデータにした場合とした場合 0.0973 と小さくなっている．これは，差をとったことにより，症例間の変動が軽減されたことによる．

症例を固定効果としても変量効果としてもプラセボと T 薬の差の差の *SE* は 0.0883 と同じで，信頼区間の 90%下限も 0.4930 と同じであるが，表 3 の元データで場合の 0.5278 よりも信頼区間の 90%下限が小さくなり，投与前値からの差による解析のメリットが見いだせない．

3.3　元データでの解析か投与前値からの差のデータでの解析か

これらの結果を踏まえて，元データ，投与前値からの差のデータ，どちらの解析結果を用いたらよいのであろうか．あるいは併記するのがよいのだろうか．投与前データの総平均に比べ群平均が高めならば増加量が抑えられ，群平均が小さめならば増加量は多めとなり，元データと差のデータでの解析結果が異なってしまう．これは，測定値がある範囲に限定されているような場合に，ある症例の投与前値が高目に出たとすれれば，その次の測定ではそれ以上になる確率は低くなり，その症例の真の平均に近づくことになる．

「平均への回帰」現象は，ICH-E14 でも，3.2 節の 中心傾向（central tendency）で説明されているが，投与前値を共変量とする共分散分析については言及されていない．投与前値の群平均が完全に一致していれば，「平均への回帰」現象の影響は受けないが，わずかでも異なると，「平均への回帰」現象の影響から逃れない．この結果として，元データでの結果と投与前からの差での結果が微妙にことなり，どちらか一方を使うと結果に対し，都合の良い方を使ったのではないかと疑われ，併記すれば，どちらの結果で判定することが望ましいのかと，詰問されることになりかねない．

3.4　投与前値を共変量とした解析

並行群間比較試験の場合には，投与前値を共変量とすることで，元データでも差のデータでも結果が一致することが知られているが，クロスオーバー法の場合でも一致するのであろうか．

図 3 に示したように投与前と 1 時間目の各薬剤での相関係数は，0.6 以上の相関となっている．

クロスオーバー法の場合には，投与前値を共変量とする場合には，各群の相関より，各症例内の相関が関与するようにも思われる，各症例内には3薬剤分のデータしかないが，1時間目のデータについてそれぞれの平均値とSDで基準化したデータについて 図 4 に散布図を描き，50%の確率楕円を上書きした結果を 図 4 に示す.

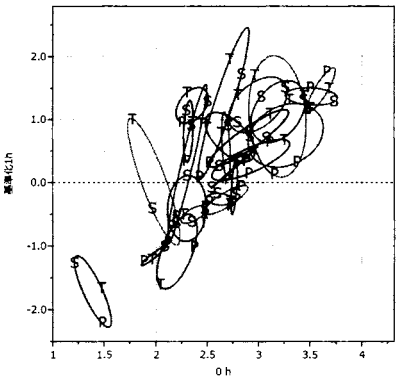


図 4　症例ごとの投与前値と1時間目の相関関係

投与前と基準化した1時間目の確率楕円から，症例内での3ポイントの相関は，正の場合もあるが，負となる場合もあり，また無相関の場合もあるが，全体的には正の相関構造が示唆される．これらの相関構造を念頭にしつつ，投与前値を共変量とした解析を行い，結果の解釈の参考とする.

症例を変量効果とし，1時間目の元データ，および投与前からの差のデータについての結果を 表 6 に示す．プラセボ群の調整平均（LSMEAN）は，2.8252で，その15%増は3.2490であり，表 3 の症例を変量効果とみなした場合の *SE* が0.1376 であるのに対し，投与前値を共変量とすることにより，0.0988 と大幅に減少し，S 薬の信頼区間の 90%下限は3.5225 とかなり上回り，T 薬は3.3076 とゆとりを持って上回るようになった.

投与前からの差について，もちろん調整平均は異なるが，*SE* は完全に一致し，クロスオーバー法であっても，元データと差のデータでも同じ結果が得られることが確認された．ただし，残念なことに薬剤間の差の *SE* は，0.0870 から 0.0850 へとわずかな減少にとどまっている．この原因は，

表 6　投与前値を共変量，症例を変量効果とした場合の信頼区間の 90%下限値

		前値を共変量：元データ			前値を共変量：差のデータ				
薬剤		調整平均	*SE*	*L* 90%	差(調整)	*SE*			P群の15%増
P		2.8252	0.0988	2.6585	0.1759	0.0988			
S		3.6892	0.0987	3.5225	1.0399	0.0987			3.2490
T		3.4743	0.0988	3.3076	0.8250	0.0988			3.2490
薬剤	薬剤	差	*SE*	*L* 90%	差の差	*SE*	*L* 90%	*p* 値	P群の15%
S	P	0.8640	0.0849	0.7213	0.8640	0.0849	0.7213	<.0001	0.4238
T	P	0.6491	0.0850	0.5063	0.6491	0.0850	0.5063	<.0001	0.4238
S	T	0.2149	0.0850	0.0721	0.2149	0.0850	0.0721	0.0151	

クロスオーバー法なので「症例」がモデルに含まれており，薬剤群間の差とした場合に各症例内で

の投与前値の共変量としての調整の役割の寄与がほとんどなくなってしまったと解される.

4. すべての測定時点を用いた解析

4.1 分割実験とみなした解析法の応用

並行群間試験の場合には，すべての測定時点を用いて解析を行うことで症例内の変動を抑えるため 1 時間目についての薬剤群での比較を行なう際に推定精度が向上することが期待される．クロスオーバー法の場合についても，同様に推定精度の向上があるのだろうか.

症例，薬剤，時点の 3 因子をランダム化の順序を考えた分割実験として考える．各症例は互いに独立していて，その中で 3 薬剤がランダムに割り付けられ，その中で 1，2,...，8 時間目のデータがランダムに測定されたとみなす．表 7 に 3 因子交互作用まで平方和を分解した結果を示す．症例は変量効果，薬剤は固定効果，薬剤×症例は変量効果で 1 次誤差，時間は固定効果であり，時間×症例，時間×薬剤×症例は，症例が変量効果なので変量効果となりも，同じ 2 次要因内なので合わせて 2 次誤差とする.

表 7　分割実験とみなした場合の平方和

	要因	自由度	平方和	平均平方	役割
ブロック	patient	23	223.97	9.7378	変量効果
1次要因	drug	2	25.78	12.8913	固定効果
	drug×patient	46	23.44	0.5096	変量：1次誤差
2次要因	time	7	17.17	2.4529	固定効果
	time×patient	161	12.21	0.0759	変量：2次誤差
	time×drug	14	6.28	0.4486	固定効果
	time×drug×patient	322	18.28	0.0568	変量：2次誤差
	全体	575	327.14		

表 7 を組替えて 表 8 に分散分析表としてまとめ直し，平均平方の構造から，分散成分計算した結果を示す．ここに示した分散分析表は，1，2,...，8 時間の測定が完全にランダム化されたとみなした解析であり，そのために 2 次誤差の自由度が 483 とインフレーションを起こし，時間×薬剤の F 検定が有意になりやすいとの批判があり，そのために自由度の補正が定式化されているが，ここでは言及しない.

表 8　分割実験とみなした場合の分散分析表

	要因	自由度	平方和	平均平方	F：1次	F：2次	平均平方の構造	分散成分
ブロック	patient	23	223.97	9.7378	19.11		$\sigma_e^2+24\sigma_{patient}^2+8\sigma_{drug\times patient}^2$	0.3845
1次要因	drug	2	25.78	12.8913	25.30		$\sigma_e^2+192\sigma_{drug}^2+24\sigma_{time\times drug}^2+8\sigma_{drug\times patient}^2$	
	drug×patient	46	23.44	0.5096	1	19.11	$\sigma_e^2+8\sigma_{drug\times patient}^2$	0.0558
2次要因	time	7	17.17	2.4529		38.86	$\sigma_e^2+72\sigma_{time}^2+24\sigma_{time\times drug}^2$	
	time×drug	14	6.28	0.4486		7.11	$\sigma_e^2+24\sigma_{time*drug}^2$	
	誤差	483	30.49	0.0631		1	σ_e^2	0.0631
	全体	575	327.14					

批判にさらされている分散分析表をあえて持ち出したのは，平均平方の構造から分散成分の推定

ができるからである．投与後の全時点を用いることにより，1 時間目のデータのみで推定した分散成分よりも，1 時間目の薬剤間の平均値に関して安定した推定値を用いることが可能となる．分散成分の推定においては，自由度のインフレーションは，平均平方の構造に示したように除去されことが確認される．

分割実験型の分散分析を行う統計ソフト（GLM プロシジャ，JMP/EMS タイプの適用）の致命的な欠陥は，分散分析表における自由度のインフレーションよりも，各種の水準間の比較にある．時間×薬剤の 1 時間目の推定値に対して薬剤群の推定平均の推定を行った時に起きる．表 9 に示すように，GLM プロシジャの RANDOM ステートメントで症例，薬剤×症例を変量と指定しても，平均平方の構造，分散分析表の F 検定は適切に対応するが，薬剤群の推定平均および差の推定平均も，すべて 2 次誤差から次のように SE が推定されているために，常に過大評価を招く．

1 時間目の薬剤群の平均 SE：　$SE_{1h,drug} = \sqrt{\hat{\sigma}_e^2 / 24} = \sqrt{0.0631/24} = 0.0513$

1 時間目の薬剤群間差の平均 SE：　$SE_{1h,drug(diff)} = \sqrt{2\hat{\sigma}_e^2 / 24} = \sqrt{2 \times 0.0631/24} = 0.0725$

症例，症例×薬剤を変量効果とした MIXED プロシジャ（JMP の REML 指定）の場合には，

1 時間目の薬剤群の均 SE：

$$SE_{1h,drug} = \sqrt{(\hat{\sigma}^2_{patient} + \hat{\sigma}^2_{patient*drug} + \hat{\sigma}^2_e)/24} = \sqrt{(0.3845 + 0.0558 + 0.0631)/24} = 0.1448$$

1 時間目の薬剤群間差の平均 SE：

$$SE_{1h,drug(diff)} = \sqrt{2(\hat{\sigma}^2_{patient*drug} + \hat{\sigma}^2_e)/24} = \sqrt{(0.0558 + 0.0631)/24} = 0.0996$$

のように，変量効果の分散成分が含まれていて，それぞれの SE は大きく推定されている．MIXED プロシジャの場合には，推定された分散成分を合成した SE を算出し，適切な推定値となる．

表 9　分割実験とみなした場合の 1 時間目の水準間の信頼区間

		1時間目	固定(GLM)		変量 (MIXED)		
薬剤		平均	*SE*	*L*90%	*SE*	*L*90%	P群の15%増
P		2.8150	0.0513	2.7305	0.1448	2.5703	
S		3.6850	0.0513	3.6005	0.1448	3.4403	3.2373
T		3.4888	0.0513	3.4042	0.1448	3.2441	3.2373
薬剤	薬剤	差	*SE*	*L*90%	SE	*L*90%	P群の15%
S	P	0.8700	0.0725	0.7505	0.0996	0.7052	0.4223
T	P	0.6738	0.0725	0.5542	0.0996	0.5090	0.4223
S	T	0.1963	0.0725	0.0767	0.0996	0.0315	

4.2　時点間の相関構造

MIXED プロシジャを用いた経時データの解析では，各種の時点間の相関構造を設定できるようになっていて，Littell らは，元データでの相関構造に対し，どのような相関構造が適合するか詳しく示し，図 5 に示すように投与後の各自時点間の相関は，時点が離れるにつれて大きくなるが，自己回帰型（AR1 タイプ）ほどではないと考察している．また，AIC を用いた相関構造の選択では，相関構造を特定しない UN タイプが優れているが，BIC の観点からで貧弱な選択だと述べ，生物学

的な観点からの検討が必要性を示唆している．

時点間の相関構造の選択は，多くの実験研究の裏付けが必要であり，また一定間隔で得られた経時データでなければ適用が困難であり，また解析しようとしている実験データそのものから特定できるものではない．ここでは，（0h，1h，2h，4h，8h）などのように測定間隔が異なる場合でも適用できる時点間の相関構造を平均的な相関としたCSタイプを用いることにする．

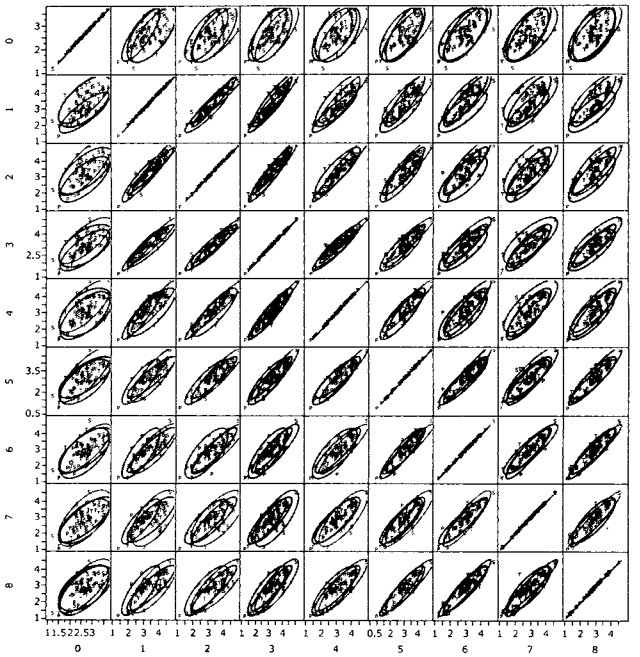


図 5　時点間の散布図と薬剤群ごとの90%確率楕円

4.3　投与前値を共変量とした解析

投与前値を共変量として変量効果モデルに組み込むことにより，1時間目の分散成分が小さくなることが期待される．表 10 に示すように，$\hat{\sigma}^2_{patient}$ は，0.3845 から 0.1669 と大幅に減少したが，$\hat{\sigma}^2_{drug\times patient}$ および $\hat{\sigma}^2_e$ の減少はほとんどなかった．

表 11 に，投与前値を共変量とした場合の1時間目における各薬剤の推定値，および薬剤群間の差の推定値と90%の信頼区間を示す．結果は，表 6 で示した1時間目の時点を用いた共分散分析とほぼ同程度の結果であり，労多くして功少なしであった．

表 10　前値を共変量とした場合の分散成分

	前値を含まず	前値を共変量	
変量効果	分散成分	分散成分	変化
patient	0.3845	0.1669	大幅減少
drug×patient	0.0558	0.0557	ほとんど変わらず
残差	0.0631	0.0631	変わらず

表 11　前値を共変量とした場合の1時間目の推定値と信頼区間

		前値を含まず（表 9 再掲）			前値を共変量			
薬剤		平均	SE	*L*90%	調整平均	*SE*	*L*90%	P群の15%増
P		2.8150	0.1448	2.5703	2.8241	0.1091	2.6410	
S		3.6850	0.1448	3.4403	3.6888	0.1091	3.5056	3.2373
T		3.4888	0.1448	3.2441	3.4758	0.1091	3.2926	3.2173
薬剤	薬剤	差	差の*SE*	*L*90%	差	差のSE	*L*90%	P群の15%
S	P	0.8700	0.0996	0.7052	0.8646	0.0995	0.6999	
T	P	0.6738	0.0996	0.5090	0.6517	0.0996	0.4868	0.4223
S	T	0.1963	0.0996	0.0315	0.2129	0.0996	0.0481	0.4223

5. 考察

TQ/TQ c 試験の，大動物を用いたテレメトリー試験ではクロスオーバー法での実験が定着しているので，元データが公開され，また各種の解析結果についても提示されている Littell らの 1 秒あたりの呼吸量 FEV1 のデータを用いて，信頼区間方式による判定を行うために，どのような統計解析が適切であるか検討した．

ICH-E14 で提示されている信頼区間方式では，時間を一致させた被験薬群とプラセボ群の平均の比較でる．並行群間比較の場合には，症例に関する変動が信頼区間に入り込み，信頼区間が広がる原因となる．さらにプラセボ群と被験薬群の平均の差の信頼区間は $SE_{群間差}=\sqrt{2\hat{\sigma}^2_{e(patient)}/n}$ となり，被験薬群の例数を 2 倍増やしても信頼区間の幅の $\sqrt{2}$ 分の 1 にしかならない．

並行群間試験の場合は，表 4 に示すように誤差分散に症例間の変動が含まれ *SE* が増大し，プラセボと T 薬の平均値の差の *SE* にも入り込んで，信頼区間の 90%下限値を押し広げる原因となる．クロスオーバー法の場合は，表 3 に示したように症例間の変動を誤差変動から分離することにより誤差分散を大幅に減少することができる．

並行群間試験とクロスオーバー法による同時点の平均値の差の信頼区間の算出方法に明らかな違いがあり，圧倒的に並行群間比較が不利である．これは，平均値の差の信頼区間を前提にしているためであり，公平な判定を行うために判定基準の明確化が必要と思われる．次のような判定基準とすることにより，試験法による差異がなくなる．

判定 1．並行群間試験でもクロスオーバー試験でも時間を一致させた被験薬群の片側 95%の信頼区間が，プラセボ群の平均値の 10 ms 増しの限界値を下回る場合に陰性とする（新たに処方される集団に対する評価）

表 3 に示した症例を変量効果とした場合に薬剤 T の信頼区間の下限は 3.2553，プラセボ P の 15%増増しは 3.2373 と，0.0181 上回っており薬効が認められる．ここで計算されている信頼区間は，表 4 に示した並行群間試験とみなして解析した結果と同様であり，試験法による判定に差異が生じにくい．しかし，この判定基準だけでは，苦労してクロスオーバー法で行った苦労が報われない．表 3 に示したクロスオーバー法での両群間の差の信頼区間の下限は，0.5278，プラセボ P の 15%増しは 0.4223 と，0.1055 とゆとりをもって上回っており，明確な薬効が認められる．しかしながら，この判定は同一症例の中で比較となり，新たに試験薬が処方される集団に対する平均的な増加を評価しているわけではない．したがって，2 つの判定を併記し，考察することを薦めたい．

判定 2. クロスオーバー法の場合，時間を一致させた被験薬とプラセボ群の差の片側 95%

の信頼区間が，10 ms を下回る場合に陰性とする（個々の症例に対する評価）

投与前値を共変量とすることで，クロスオーバー法の場合でも，ある時間の元データでの結果と投与前値からの差のデータでも，表 6 に示したように投与群の *SE*，および群間の *SE* が完全に一致することが確認され，これにより，元データか差のデータかでの結果の不一致からくる悩ましい問題の解決となる．ICH-E14 では，中心傾向の解析のため投与前値からの差のデータについて時間を一致させたプラセボと被験薬の平均値の差に対して多面的に検討するように求めているが，投与前のベースライン値を共変量とした解析を標準的に使うべきである．

クロスオーバー法で，時期ごとの投与前値を共変量として用いない場合に対する相対効率を Yan（2011）が次のように報告している．時期数が 2 の場合に相対効率は高いが，時期数が 3 で，薬剤数が 3 の場合には時点間の相関が高ければ相対効率は上がるが，高々数パーセントでしかないことが示されている．今回の群間差の *SE* が元データの場合 0.5278，差のデータの場合 0.4930，共変量とした場合 0.5063 であり，Yan が示した結果と整合している．

クロスオーバー試験では症例を固定効果とするか変量効果とするかによって薬剤群の平均値に対する *SE* が大きく異なり，固定効果とした場合には，同じ被験者に再度実験をした場合の評価となり，一般化可能性を考えない判定になるので，推奨することができない．また症例を変量効果と指定しても REML 法での計算を行うことが必須である．

測定時点を固定効果としてモデルに取り込む，いわゆる反復測定共分散分析の適用は，クロスオーバー法での Littell らの実験データの解析では，分析精度の向上が確認できなかった．並行群間試験の場合には，複数時点のデータを取り扱うことによって分散の安定化が図れて分析精度の向上が期待できるのであるが，クロスオーバー法の場合には，同じ症例に対する同時刻で繰り返し測定が解析モデルの中で扱われており，経時データとしての解析において，分析性度の向上を見いだせなかった．

今後，さらに検討を要するが，時点を含めることにより複雑な解析を行うよりも，時点ごとあるいは，数時点の平均値による解析を時点ごとに繰り返し，信頼区間による総合的な判定を行うことを推奨する．

文 献

Yan Z.（1997），The impact of baseline covariates on the efficiency of statistical analyses of crossoverdesigns. Statstics in Medicine. 32: 956-963.

Littll, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (2006) *SAS System for Mixed Models 2ed.* SAS Institute.

渡橋靖（2009），QT 延長をいかに判定すべきか：ICH E14 に準拠した臨床試験デザインと統計解析法，日薬理誌（Folia Pharmacol. Jpn.）133，14-8.

高橋行雄（1996），各種の分割実験および経時測定データの解析，SAS ユーザ会論文集，263-86.

高橋行雄（2009），薬理学研究における経時データ解析の考え方 —血圧降下試験事例による解説—，日薬理誌（Folia Pharmacol. Jpn.）133，325-31.

高橋行雄(2010)，経時データに対する投与前値を考慮した解析モデルの比較検討，SAS ユーザ会論文集，45-54.

付録：データリスト

	P(p)			S(c)			T(a)		
patient	0 h	1 h	d(1-0 h)	0 h	1 h	d(1-0 h)	0 h	1 h	d(1-0 h)
201	2.14	2.36	0.22	2.30	3.41	1.11	2.46	2.68	0.22
202	3.37	3.03	-0.34	2.91	3.92	1.01	3.50	3.95	0.45
203	1.88	1.99	0.11	2.08	2.52	0.44	1.96	2.28	0.32
204	3.10	3.24	0.14	3.02	4.43	1.41	3.44	4.08	0.64
205	2.91	3.35	0.44	3.26	4.55	1.29	2.80	4.09	1.29
206	2.29	3.04	0.75	2.29	4.25	1.96	2.36	3.79	1.43
207	2.20	2.46	0.26	1.96	3.00	1.04	1.77	3.82	2.05
208	2.70	2.85	0.15	2.70	4.06	1.36	2.64	3.67	1.03
209	2.25	3.45	1.20	2.50	4.37	1.87	2.30	4.12	1.82
210	2.48	2.56	0.08	2.35	2.83	0.48	2.27	2.77	0.50
211	2.12	2.19	0.07	2.34	4.06	1.72	2.44	3.77	1.33
212	2.37	2.14	-0.23	2.20	2.82	0.62	2.04	2.00	-0.04
214	2.73	2.57	-0.16	2.78	3.18	0.40	2.77	3.36	0.59
215	3.15	2.90	-0.25	3.43	4.39	0.96	2.96	4.31	1.35
216	2.52	3.02	0.50	3.07	3.90	0.83	3.11	3.88	0.77
217	1.48	1.35	-0.13	1.21	2.31	1.10	1.47	1.97	0.50
218	2.52	2.61	0.09	2.60	3.19	0.59	2.73	2.91	0.18
219	2.90	2.91	0.01	2.61	3.54	0.93	3.25	3.59	0.34
220	2.83	2.78	-0.05	2.48	2.99	0.51	2.73	2.88	0.15
221	3.50	3.81	0.31	3.73	4.37	0.64	3.30	4.04	0.74
222	2.86	3.06	0.20	2.54	3.26	0.72	2.85	3.38	0.53
223	2.42	2.87	0.45	2.83	4.72	1.89	2.72	4.49	1.77
224	3.66	3.98	0.32	3.47	4.27	0.80	3.68	4.17	0.49
232	2.88	3.04	0.16	2.79	4.10	1.31	2.49	3.73	1.24
平均	2.636	2.815	0.179	2.644	3.685	1.041	2.668	3.489	0.820
SD	0.518	0.577	0.337	0.546	0.707	0.477	0.554	0.728	0.584

医療・臨床研究分野におけるCDISC標準規格群への取り組み

大津　洋
順天堂大学大学院医学研究科
先導的がん医療開発研究センター

木内　貴弘
東京大学大学院医学系研究科公共健康医学専攻
医療コミュニケーション学・UMIN センター

Efforts to CDISC standards in the medical and clinical research field in Japan

Hiroshi Ohtsu, M.Sc.
Leading Center for the development and research of Cancer medicine,
Juntendo University

Prof. Takahiro Kiuchi
Department of Health Communication,
School of Public Health, The University of Tokyo

要旨

医療・臨床研究分野での CDISC 標準規格群がどのように紹介され、また、現状がどのような形になっているのか、大学病院医療情報ネットワーク研究センター（UMIN）、東京大学での取り組みを中心に取り上げる。臨床研究における標準化されたデータ規格の取扱いは、もはや治験だけでなく広く活用されるべきであり、それに向けて取り組みについても紹介する。

キーワード：UMIN , CDISC, ODM, SDTM , ADaM, Academic research

医療・臨床研究分野での CDISC 標準規格群

医療・臨床研究領域で CDISC 標準規格群に触れ始めたのは、平成 16 年度厚生労働科学研究費補助金厚生労働科学特別研究事業「次世代医療機器研究・開発・商業科促進のための薬事承認のあり方に関する研究（主任研究者：砂川賢治）」での報告書が先行しており、また、平成 17 年度　日本医師会治験推進センター治験推進事業「治験の IT 化の現状と課題（主任研究者：木内貴弘）」の研究の一部として、米国への訪問と詳細な報告書にまとめられている。平成 17 年の報告については、 [古川浩之、他, 2006]を始め、 [木内貴弘、大津洋, 2008] にて CDISC 標準規格群を利用することによる統計解析業務への影響の考察など、いくつかのレポートとして発表されている。しかし、当時の状況では、大きなインパクトにはなりえなかった。
しかしながら、我々は、CDISC 標準規格群の影響力の大きさを見越し、継続的に海外との情報交流や、「新たな治験活性化 5 か年計画（平成 19 年 4 月実施）の中で、「CDISC 標準」という言葉を治験電子化のキーワードのひとつとして提案し、それを文言に含めることができた。この経緯については、 [木内貴弘, 2013]に詳細に記載しているので、そちらを参照するとよい。

UMIN センターを中心とする取組- 現状

近年、CDISC 標準規格群について検討・利用を開始した場合、対 PMDA としての規格群、CDISC SDTM/ADaM に注目し、その規格についてのみ調査することが多い。あくまでも CDISC SDTM/ADaM は、臨床試験を行った後の成果物のひとつのデータベースであり、統計解析を実施するものとしては充足するだろうが、データマネージャー、プロジェクト担当者、システム担当者からすると、CDISC SDTM/ADaM ベースだけではあくまでもデータ層のみの発想であり、通信層や定義層をカバーできていないという問題点がある。
その点、我々は、行政当局からの通知が行われる前からの活動を続けており、承認申請がある・なしに関わらず、CDSIC 標準規格群を調査することが可能であった。従って、情報バイアスが入ることなく調査・検討ができているといえよう。
情報の流通・情報の格納　の観点から、出来るだけ正確にシステム化を実施するため、通信規約である ODM とデータセンター(UMIN センター)を結ぶという手法を取ってきた。UMIN は、日本の研究領域では最大級のデータセンター(UMIN INDICE)を提供しているが、他のシステムと UMIN INDICE を CDISC ODM で通信するための基本的な規約である" UMIN INDICE Lower level data communication protocol of for CDISC ODM" を 2013 年 7 月に公開しており、医療機関側からでも、他の e-CRF からでも、システム対応が可能な状態にしていることが特徴的である。
また、CDISC 標準規格群を「治験」以外でも使える基盤として、平成 23 年度に「CDISC 標準を活用した死体検案書の施設別及び全国集計データベースの構築」の実施を行い、CDISC ODM でデータベースと通信を行うクライアントソフトの構築をし、実運用に至っている。

医療・臨床研究分野でどう使うべきか？

医療・臨床研究分野では、いくつかの大きな変動があった。研究者主導治験が増加傾向であること、また、昨年来、臨床研究における研究不正と思われる事案が複数発覚したことである。今後の臨床試験においては、行政当局に申請のあるなしに関わらず、医療機関からのデータ提供が正しいことも当然であるが、その後のデータの流れ（データの入手から解析結果の算出まで）を、これまで以上に明確にしておかなくてはならない状況になりつつある。以前は、EDC での（いわゆる）コンピュータ上でのロジカルチェックを厳密にして、入力を制御する方向であったが、そうではなく、入力はできる限り早く行われ、変更がきちんと記録されていること、どの情報がどういうプロセスを経て、結果として示されているのか、という点である。
その点では、CDISC 標準規格群をフルに活用できれば、ある程度のプロセスを標準化し、最適化に向かわせる可能性を秘めている。幸いにして、CDISC 標準規格群を使った医療機関等からのデータ収集の面では、現状、日本が世界をリードしている。それは、前述した UMIN もそうであるが、それ以外の研究機関や団体においても、いくつか CDISC 標準規格群を使ったシステム開発を行ってきている経緯が、欧米の企業と FDA がモデルケースを立ち上げ、そこから実運用に持っていくことと逆のアプローチであるからである。
この日本流のアプローチは、実運用例を積み上げて、治験におけるデータ収集に高めていくという点では有効な手であると考えられるが、その一方で、規格への誤解などで、同じ CDISC 標準群と言いながら、亜流が生まれる危惧も拭い去れない。統制が取れる欧米流も一理ある方向ではある。お互いに利活用していくことで、"One-standard" な規格になるように、実務者として取り組むべきである。
さて、今後、標準化した規格を多くの人に正しく理解して、使ってもらうためには、広く教育活動が必須で

あると言えよう。

我々は、これまでの公的研究費を用いた研究の成果や経験をもとに、CDISC 標準規格群についての理解を含める講習会を実施してきた。UMIN 主催のものとしては、2008 年と 2014 年に実施したが、社会人を中心に多くの参加者があり、関心の高さが伺われた。また、SAS においても、"Implementing CDISC Using SAS:An End-to-End Guide (SAS Press)"を基とした CDSIC 標準規格群における SDTM/ADaM の実装について、本の内容と、近年の動向を踏まえて講義を複数回実施している。

関心のある人達への教育活動もさることながら、将来の臨床試験の姿を考えると、データ構造の標準化が一般化することが想定される。現状は産官学の先端を目指すものとして CDISC 標準規格群への理解は必要な知識であるという認識より、本年より東京大学　医学系研究科公共医学専攻にて「医学研究と CDISC 標準」を単位認定科目として開講することになっている。

最後に若手の統計担当者・学生に対して、CDISC ADaM を理解し使ってもらうという点では、若干対応が遅れている。この点については、東京理科大の佐野先生のグループの活動に期待したい。

参考文献

古川浩之、他. (2006). 臨床試験データの電子的伝達の標準化. 月刊薬事, 125(1769) - 134(1778).

木内貴弘. (2013). 日本のアカデミアにおける CDISC 標準への取り組み. Jpn Pharmacol Ther Vol.41 suppl, 13-18.

木内貴弘、大津洋. (2008). CDISC 標準の現状と今後及び臨床研究データ管理・統計解析への影響. Proc Soc Clin Biostat Res. 28, 39-49.

SASとExcelを用いたCDISC ADaM標準における作業効率化の試み

高浪 洋平
武田薬品工業株式会社医薬開発本部日本開発センター
クリニカルデータサイエンス部統計グループ

Approach to Reduction in Workload of CDISC ADaM Standard Using SAS and Excel

Yohei Takanami
Takeda Pharmaceutical Company, Ltd.

要旨

医薬品医療機器総合機構（PMDA）より H28 年度以降の医薬品承認申請時における臨床試験電子データ提出義務化が発表され（詳細は http://www.pmda.go.jp/operations/shonin/info/iyaku/jisedai.html 参照），製薬企業側は，CDISC に準拠したプロトコルの作成をはじめ，今後様々な業務において対応に追われることが予想される．特に，製薬企業の DM・統計解析・SAS プログラマーの担当者は，臨床試験データ標準である Study Data Tabulation Model（SDTM），解析用データ標準である Analysis Data Model（ADaM）形式に従ったデータセット及び関連する文書（Define.xml や解析関連の SAS プログラム等）の作成が求められ，今後の工数の増大は避けられない．本発表では，SAS プログラムと Excel Metadata 及び Excel VBA の極めて基本的な機能のみを用いて，ADaM の Define.xml の作成や解析及びバリデーションの実施における作業効率化の試みを紹介する．

キーワード：CDISC，ADaM，Excel，Metadata，VBA，Define.xml，OpenCDISC

1 ADaM データセットと Metadata

将来的に CDISC 標準に準拠した臨床試験電子データを FDA や PMDA 等の規制当局に提出する際に ADaM データセットを作成する場合，データセットとともに，ADaM V2.1 に定義されている各 Metadata を最終的に Define.xml として作成して提出することが想定される（FDA（2014），PMDA（2014））．Metadata や Define.xml の作成方法はいくつか提案されているが，本稿では，Jack Shostak ら（2012）で紹介されている方法をもとに，Excel 形式の Metadata を作成する．また，次章ではその Excel Metadata の情報をもとに SAS プログラムによる Define.xml の作成方法を提案する．

1.1 ADaM データセット

本稿では，表 1.1 に示す架空の臨床試験から得られたデータを想定し，ADaM データセットとして，ADSL，ADAE 及び BDS 構造に従ったデータセットを用意する．

項目	内容
相	第 III 相検証試験（試験名：PROD-XXX/STUDY-XXX）
対象	疾患 A を有する被験者
試験デザイン	2 群（被験薬群（Drug A）・対照薬群（Drug B））並行群間無作為化比較試験 1 日 1 回経口投与（ダブルダミー）
目的	被験薬と対照薬の疾患 A の治癒率を比較する.
主要評価項目	投与 8 週後の疾患 A の治癒率の比較（対照薬群との治癒率の差を検証する）
その他の評価項目	有害事象等

表 1.1 想定する臨床試験の概要

ADaM データセットの概要を図 1.1 に示す.

データセット：内容（一部抜粋）

ADSL：
被験者情報を格納したデータセット（1 症例 1 レコード）

STUDYID	USUBJID	SUBJID	SITEID	AGE	AGEU
PROD-XXX/STUDY-XXX	PROD-XXX/STUDY-XXX/9001001	9001001	9001	70	YEARS
PROD-XXX/STUDY-XXX	PROD-XXX/STUDY-XXX/9001002	9001002	9001	63	YEARS
PROD-XXX/STUDY-XXX	PROD-XXX/STUDY-XXX/9001003	9001003	9001	56	YEARS
PROD-XXX/STUDY-XXX	PROD-XXX/STUDY-XXX/9001004	9001004	9001	62	YEARS
PROD-XXX/STUDY-XXX	PROD-XXX/STUDY-XXX/9001005	9001005	9001	78	YEARS
PROD-XXX/STUDY-XXX	PROD-XXX/STUDY-XXX/9001006	9001006	9001	61	YEARS
PROD-XXX/STUDY-XXX	PROD-XXX/STUDY-XXX/9001007	9001007	9001	38	YEARS
PROD-XXX/STUDY-XXX	PROD-XXX/STUDY-XXX/9001008	9001008	9001	38	YEARS
PROD-XXX/STUDY-XXX	PROD-XXX/STUDY-XXX/9001009	9001009	9001	76	YEARS
PROD-XXX/STUDY-XXX	PROD-XXX/STUDY-XXX/9002001	9002001	9002	67	YEARS
PROD-XXX/STUDY-XXX	PROD-XXX/STUDY-XXX/9002002	9002002	9002	37	YEARS

ADAE：
有害事象の情報を格納したデータセット（1 事象 1 レコード）

AETERM	AEDECOD	AEPTCD	AEBODSYS	AEBDSYCD	AELLT
Cystitis acute	Cystitis	10011781	Infections and infestations	10021881	Cystitis acute
Diarrhoea	Diarrhoea	10012735	Gastrointestinal disorders	10017947	Diarrhoea
Alopecia areata	Alopecia areata	10001761	Skin and subcutaneous tissue disorders	10040785	Alopecia areata
Mallory-Weiss syndrome	Mallory-Weiss syndrome	10026712	Gastrointestinal disorders	10017947	Mallory-Weiss syndrome
Common cold	Nasopharyngitis	10028810	Infections and infestations	10021881	Common cold
Acute abdomen	Acute abdomen	10000647	Gastrointestinal disorders	10017947	Acute abdomen
Common cold	Nasopharyngitis	10028810	Infections and infestations	10021881	Common cold
Gastric polyps	Gastric polyps	10017817	Gastrointestinal disorders	10017947	Gastric polyps
Headache	Headache	10019211	Nervous system disorders	10029205	Headache
Constipation	Constipation	10010774	Gastrointestinal disorders	10017947	Constipation
Paronychia	Paronychia	10034016	Infections and infestations	10021881	Paronychia
Epigastric discomfort	Epigastric discomfort	10053155	Gastrointestinal disorders	10017947	Epigastric discomfort
Pharyngeal erosion	Pharyngeal erosion	10052773	Respiratory, thoracic and mediastinal disorders	10038738	Pharyngeal erosion

ADEF：
BDS 構造で，疾患 A の治癒・未治癒判定を格納したデータセット（1 時点・1 パラメータで 1 レコード）
※パラメータは一つのみ

ADY	ADYL	PARAM	PARAMCD	PARAMN	AVAL	AVALC	AVISIT	AVISITN	ANL01FL
13	-16	Healing of Disease A	HEAL	1	2	Unhealed	Week 2	2	Y
30	1	Healing of Disease A	HEAL	1	1	Healed	Week 4	3	Y
30	1	Healing of Disease A	HEAL	1	1	Healed	Week 8	4	Y
16	-13	Healing of Disease A	HEAL	1	2	Unhealed	Week 2	2	Y
30	1	Healing of Disease A	HEAL	1	1	Healed	Week 4	3	Y
30	1	Healing of Disease A	HEAL	1	1	Healed	Week 8	4	Y
14	-14	Healing of Disease A	HEAL	1	2	Unhealed	Week 2	2	Y
29	1	Healing of Disease A	HEAL	1	1	Healed	Week 4	3	Y
29	1	Healing of Disease A	HEAL	1	1	Healed	Week 8	4	Y
15	1	Healing of Disease A	HEAL	1	1	Healed	Week 2	2	Y
15	1	Healing of Disease A	HEAL	1	1	Healed	Week 4	3	Y
15	1	Healing of Disease A	HEAL	1	1	Healed	Week 8	4	Y

図 1.1 ADaM データセットの概要

1.2 Excel 形式の ADaM Metadata の作成

ADaM では，以下の Metadata が定義されており，最終的にこれらの情報を Define.xml に格納する必要がある.

種類	概要
Analysis Dataset Metadata	解析用データセットの一覧と概要が記載されたメタデータ
Analysis Variable Metadata	各解析用データセットに含まれる変数のラベル，コードリスト，導出方法等の情報が記載されたメタデータ
Analysis Parameter Value-Level Metadata	各解析パラメータの導出方法等の情報が記載されたメタデータ
Analysis Results Metadata	実施した解析の概要やレコード抽出条件，プログラム等の情報が記載されたメタデータ

表 1.2 ADaM の Metadata の概要

表 1.2 に加えて，Define.xml を作成する際にはコードリストの情報が必要なため，同様に作成して管理しておくと効率的である．本稿では，Jack Shostak ら（2012）の方法をもとに，Excel 形式の Metadata の一部として，ADaM データセットで用いられているコードリストの一覧も Metadata に含めて作成した．各シートのイメージを図 1.2 に示す．

Analysis Dataset Metadata シート：データセットの一覧

Dataset Name	Dataset Description	Dataset Locatio	Dataset Structure	Key Variables of Dataset	Class of Dataset	Documentation
ADSL	Subject disposition, demographic, and baseline characteristics	ADSL.xpt	one record per subject	USUBJID	ADSL	SAP, ADSL.sas
ADAE	Adverse Event Analysis Dataset	ADAE.xpt	one record per subject per each AE recorded in SDTM.AE domain	USUBJID, AESEQ	ADAE	ADAE.sas Dictionary used is MedDRA VXX.X
ADEF	Analysis Dataset for Target Disease	ADEF.xpt	1 record per subject, parameter	USUBJID, PARAMCD	BDS	SAP, ADEF.sas

ADSL シート：ADSL データセットの変数一覧（ADAE，ADEF も同様に作成）

Dataset Nam	Variable Name	Variable Label	Variable Type	Length	Display Forma	Codelist / Controlled Terms	Codelist Name	Origin	Source / Derivation
ADSL	STUDYID	Study Identifier	text	40	$40.			Predecessor	DM.STUDYID
ADSL	USUBJID	Unique Subject Identifier	text	40	$40.			Predecessor	DM.USUBJID
ADSL	SUBJID	Subject Identifier for the Study	text	20	$20.			Predecessor	DM.SUBJID
ADSL	SITEID	Study Site Identifier	text	20	$20.			Predecessor	DM.SITEID
ADSL	AGE	Age	integer	8	3.0			Predecessor	DM.AGE
ADSL	AGEU	Age Units	text	20	$20.	(AGEU)	AGEU	Predecessor	DM.AGEU
ADSL	SEX	Sex	text	1	$1.	M, F	SEX	Predecessor	DM.SEX
ADSL	SEXN	Sex (N)	integer	8	1.0	1=Male, 2=Female	SEXN	Assigned	1 if DM.SEX="M", 2 if DM.SEX="F"
ADSL	RACE	Race	text	20	$20.	(RACE)	RACE	Predecessor	DM.RACE
ADSL	RACEN	Race (N)	integer	8	1.0		RACEN	Assigned	1 if DM.RACE = "ASIAN"
ADSL	ARM	Description of Planned Arm	text	20	$20.	Drug A Drug B Screen Failure	ARM	Predecessor	DM.ARM
ADSL	TRT01P	Planned Treatment	text	20	$20.	Drug A, Drug B	TRT	Predecessor	DM.ARM (missing if DM.ARM="Screen Failure")
ADSL	TRT01PN	Planned Treatment (N)	integer	8	1.0	1=Drug A, 2=Drug B	TRTN	Assigned	1 if DM.ARMCD="Drug A", 2 if DM.ARMCD="Drug B"
ADSL	TRT01A	Actual Treatment	text	20	$20.	Drug A Drug B	TRT	Predecessor	DM.ACTARM (missing if DM.ACTARM="Screen Failure")
ADSL	TRT01AN	Actual Treatment (N)	integer	8	1.0	1=Drug A, 2=Drug B	TRTN	Assigned	1=Drug A, 2=Drug B

Codelist シート：コードリストの一覧

Name	CodeValue	CodeText	Data Type
SEX	F	Female	text
SEX	M	Male	text
SEXN	1	Male	integer
SEXN	2	Female	integer
AGEU	YEARS		text
ARM	Drug A	Drug A	text
ARM	Drug B	Drug B	text
ARM	Screen Failure	Screen Failure	text
TRT	Drug A	Drug A	text
TRT	Drug B	Drug B	text
TRTN	1	Drug A	integer
TRTN	2	Drug B	integer

Value List シート：解析パラメータの一覧

Name	Dataset Name	Variable Name	Variable Label	Parameter_variable	Comparator	Parameters	Variable Type	Length	Display Format
VL.ADEF.AVAL	ADEF	AVAL	Analysis Value	PARAMCD	IN	HEAL	integer	8	1.0

Analysis Results Metadata シート：解析結果メタデータ

Display Identifier	Display Name	Result Identifier	PARAM.LIST	Parameter	Analysis Variable	Reason	Dataset	Selection Criteria	Documentation	Programming Statements
Table 1	Disease A Healing Rate by Study Visit (FAS)		Healing of Disease A	HEAL	AVAL	Pre-specified in SAP	ADEF	FASFL="Y" and PARAMN = 1 and AVISITN = 4	SAP	proc format ; value _TRTPF 1 = "Drug A" 2 = "Drug B" ; run ; proc freq data=_ADaM.ADEF ; where FASFL = 'Y' and PARAMN = 1 ; table AVISIT*TRTPN*AVALC / riskdiff nocol nopc format TRTPN _TRTPF. ; run ;
Table 2	Disease A Healing Rate by Study Visit (PPS)		Healing of Disease A	HEAL	AVAL	Pre-specified in SAP	ADEF	PPROTFL="Y" and PARAMN = 1 and AVISITN = 4	SAP	proc format ; value _TRTPF 1 = "Drug A" 2 = "Drug B" ; run ; proc freq data=_ADaM.ADEF ; where PPROTFL = 'Y' and PARAMN = 1 ; table AVISIT*TRTPN*AVALC / riskdiff nocol nopc format TRTPN _TRTPF. ; run ;

図 1.2 Excel Metadata の概要（ファイル名：Run_SAS.xlsm）

次章では，これらの Excel Metadata を用いて，ADaM データセットに関連する作業の効率化を提案する．

2 Excel VBA と SAS プログラムの連携による作業効率化の試み

本章では，第 1 章で作成した ADaM データセットと Excel ファイル形式の Metadata の情報から，VBA と SAS を用いて Define.xml 及び解析帳票の作成，さらに OpenCDISC Validator（以下 OpenCDISC）によるバリデーションを実施する方法を提案する．表 2.1 に示す Excel VBA 実行用のシートを Excel Metadata のファイルに順次追加し，それらのシートから VBA を用いて SAS を実行する．VBA による SAS の実行方法及び SAS による Excel ファイルの読み込み方法についても後述する．

追加するシート	概要
Define.xml（図 2.6）	Define.xml ファイルを作成するための情報と SAS 実行ボタンを配置する．
SUMMARY（図 2.10）	被験者背景等の要約表を作成するための情報と SAS 実行ボタンを配置する．
AE（図 2.10）	有害事象の集計表を作成するための情報と SAS 実行ボタンを配置する．
Validation（図 2.13）	OpenCDISC を実行するための情報と SAS 実行ボタンを配置する．

表 2.1 SAS を実行するためのシート名

Excel Metadata に含まれる各シートと，VBA 及び SAS の実行時のイメージを以下に示す．表 2.1 の各シートから VBA を用いて SAS を実行する．

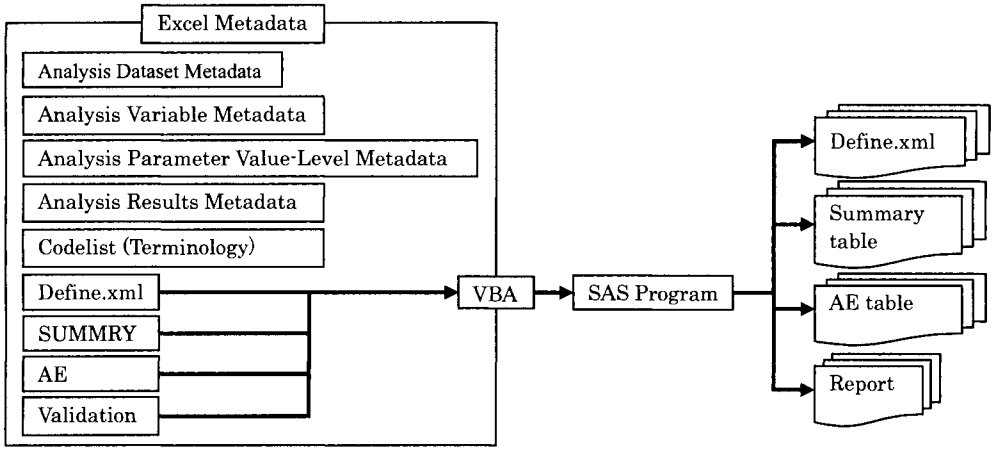


図 2.1 Excel Metadata のシートと処理フロー

2.1 フォルダ構成

本稿では，「c:¥temp¥SUGI_2014」フォルダの下に関連するファイルを全て格納して管理する．ADaM データセット，Metadata，作成されるファイル等のフォルダ構成を図 2.2 に示す．

	フォルダ名	概要
Data Define_xml Log Metadata Output Programs	Data	ADaM データセットを格納する．
	Define_xml	Define.xml ファイルを作成するプログラムと結果を格納する．
	Log	プログラム実行時のログを格納する．
	Metadata	第 1 章で作成した Excel 形式の Metadata ファイルを格納する．
	Output	解析帳票や OpenCDISC の実行結果を格納する．
	Programs	解析帳票や OpenCDISC を実行するプログラムを格納する．

図 2.2 「c:¥temp¥SUGI_2014」フォルダ内の構成

2.2 Excel VBA による SAS プログラムの実行

Excel では，VBA とフォームのボタンを組み合わせることで，SAS を実行することができる．本稿では，森岡（2013）が紹介した方法を用いて，以下のコードを「実行」ボタンに実装した．図 2.3 では，シートに配置した「実行」ボタンクリック時に SAS が起動され，%inc ステートメントで，Define.xml を作成する「Mr_Define.sas」という SAS プログラムが実行される．

実行ボタンのフォーム	VBA のコード
実行	Sub SAS_RUN3() Dim sasobj As Object Set sasobj = CreateObject("SAS.application") sasobj.Visible = False sasobj.Submit ("%inc 'C:¥Temp¥SUGI_2014¥Define_xml¥Mr_Define.sas';") sasobj.Submit ("ENDSAS;") End Sub

図 2.3 Excel から SAS を実行する VBA マクロ

2.3 SAS による Excel データの読み込み

SAS の libname ステートメントの excel エンジンでは，Excel 形式のファイルをライブラリ，各シートをデータセットとして処理を行うことができ，容易に Excel ファイルの情報を入出力できるため，非常に有用な機能である（森岡（2013），高浪ら（2012））．プログラム上では，libname ステートメントに Excel ファイルのフルパスと excel エンジンを指定し，シートを読み込む際は「'シート名$'n」と記述して，各シートをデータセットのように扱うことができる．図 2.4 では，図 1.2 に示した「Codelist」シートを読み込むプログラムと読み込まれたデータセット「_CODELIST」を示す．

```
*--- Metadata file ;
libname _META excel "C:¥temp¥SUGI_2014¥Metadata¥Run_SAS.xlsm" ;
*** Excel から Codelist シートをデータセットとして読み込む ;
data _CODELIST ;
   set _META.'Codelist$'n ;
run ;
```

データセット「_CODELIST」（一部抜粋）

Name	CodeValue	CodeText	Data Type
SEX	F	Female	text
SEX	M	Male	text
SEXN	1	Male	integer
SEXN	2	Female	integer
AGEU	YEARS		text
ARM	Drug A	Drug A	text
ARM	Drug B	Drug B	text
ARM	Screen Failure	Screen Failure	text
TRT	Drug A	Drug A	text
TRT	Drug B	Drug B	text
TRTN	1	Drug A	integer
TRTN	2	Drug B	integer

図 2.4 libname excel エンジンによる Excel ファイルの読み込み

この方法により，Excel の各シートに格納されている情報を容易に SAS のデータセットやマクロ変数として利用することが可能となる．

2.4 SAS プログラムによる Define.xml の作成

Jack Shostak ら (2012) は, Excel Metadata を SAS で読み込んで Define.xml を作成する方法を提案した. また, 高浪 (2013) は, HTML Application と SAS を用いた GUI ベースでの Define.xml 作成方法を提案したが, 本稿では Excel Metadata を使用し, VBA から SAS プログラムを実行して Define.xml を作成する方法を提案する. 「Define.xml」シートによる Define.xml 作成フローを図 2.5 に示す.

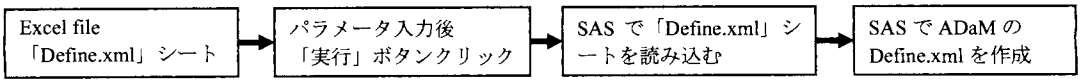


図 2.5 Define.xml 作成フロー

続いて, 「Define.xml」シートのレイアウトを以下に示す. Item 列には各パラメータの概要, Value 列には選択されたフォルダやテキスト入力及びリスト入力によって値を入力する. MVNAME 列は, SAS 実行時に格納されるマクロ変数の名前となる. Value 列に出力フォルダと必要な情報を入力し, 「実行」ボタンをクリックすると, 図 2.3 で示した VBA が実行され, 続いて Define.xml 作成 SAS プログラム「Mr_Define.sas」が Excel Metadata の情報をもとに ADaM の Define.xml を作成する. HTML ファイル及び Analysis Results Metadata の作成の有無も指定可能となっているが, ここではどちらも作成 (Yes) を指定する.

Item	Value		MVNAME
Path of Main folder ("temp" folder in this folder should exist)	C:\Temp\SUGI_2014\Define_xml files	参照	_CPATH
Study/Protocol Name	XXX/YYY		_STUDY
Creation of Define.html (MSXSL.exe should exist in the "files" folder.)	Yes		_HTML
Creation of Analysis Results Metadata	Yes		_ARM
			実行

図 2.6 Define.xml シート

実行される SAS プログラム「Mr_Define.sas」の一部を以下に示す. パラメータを Excel シートから読み込んだ後, 順次必要なタグセットを出力していくプログラムが繰り返される.

```
********** Header of the Define_ADaM.xml ********** ;
filename _H "&_CPATH.¥temp¥_header.txt" ;
data _HEADER ;
   file _H ;
   _DT = put(datetime(), E8601DT.) ;
   put '<?xml version="1.0" encoding="UTF-8"?>' ;
   put '<?xml-stylesheet type="text/xsl" href="define2-0-0_MOD.xsl"?>' ;
   put '<ODM xmlns="http://www.cdisc.org/ns/odm/v1.3"' ;
   put '     xmlns:def="http://www.cdisc.org/ns/def/v2.0"' ;
   put '     xmlns:xlink="http://www.w3.org/1999/xlink"' ;
   %if &_ARM = 1 %then put '     xmlns:adamref="http://www.cdisc.org/ns/ADaMRes/DRAFT"' ;;
   put '     ODMVersion="1.3.2"' ;
   put '     FileOID="' "&_STUDY-Define-XML_2.0.0" '"' ;
   put '     FileType="Snapshot"' ;
   put '     CreationDateTime="' _DT + (-1) '"' ;
   put '     Originator="CDISC ADaM Metadata Team">' ;
   put '  <Study OID="' "&_STUDY" '">' ;
   put '    <GlobalVariables>' ;
   put '      <StudyName>' "&_STUDY" '</StudyName>' ;
   put '      <StudyDescription>' "&_STUDY Data Definition" '</StudyDescription>' ;
   put '      <ProtocolName>' "&_STUDY" '</ProtocolName>' ;
```

```
    put '    </GlobalVariables>' ;
    put '    <MetaDataVersion OID="MDV.' "&_STUDY" '.ADaMIG.1.0.ADaM.2.1"' ;
    put '                     Name="' "&_STUDY, Data Definitions" '"' ;
    put '                     Description="' "&_STUDY, Data Definitions" '"' ;
    put '                     def:DefineVersion="2.0.0"' ;
    put '                     def:StandardName="ADaM-IG"' ;
    put '                     def:StandardVersion="1.0">' ;
run ;
```

図 2.7 Define.xml のヘッダー部分作成プログラム（抜粋）

実行結果の一部を以下に示す．本稿では，基本的に Define.xml のバージョン 2.0 のスキーマに従って作成しているが，Analysis Results Metadata のスキーマを拡張しておらず，xpt ファイル以外の外部ファイルへのリンク機能は付与されていない．

```
<?xml-stylesheet type="text/xsl" href="define2-0-0_MOD.xsl"?>
<ODM xmlns="http://www.cdisc.org/ns/odm/v1.3"
     xmlns:def="http://www.cdisc.org/ns/def/v2.0"
     xmlns:xlink="http://www.w3.org/1999/xlink"
     xmlns:adamref="http://www.cdisc.org/ns/ADaMRes/DRAFT"
     ODMVersion="1.3.2"
     FileOID="XXX/YYY-Define-XML_2.0.0"
     FileType="Snapshot"
     CreationDateTime="2014-06-23T11:40:16"
     Originator="CDISC ADaM Metadata Team">
  <Study OID="XXX/YYY">
    <GlobalVariables>
      <StudyName>XXX/YYY</StudyName>
      <StudyDescription>XXX/YYY Data Definition</StudyDescription>
      <ProtocolName>XXX/YYY</ProtocolName>
    </GlobalVariables>
    <MetaDataVersion OID="MDV.XXX/YYY.ADaMIG.1.0.ADaM.2.1"
                     Name="XXX/YYY, Data Definitions"
                     Description="XXX/YYY, Data Definitions"
                     def:DefineVersion="2.0.0"
                     def:StandardName="ADaM-IG"
                     def:StandardVersion="1.0">
```

図 2.8 作成された ADaM 用の Define.xml（抜粋）

2.5 SAS プログラムによる解析帳票の作成

Excel Metadata と VBA を使用することで，Excel シートを入力画面として必要なパラメータを入力して解析帳票を作成することができる．ここでは，Excel シートに ADaM データセットと変数を入力し，被験者背景と有害事象の要約表を作成するための「SUMMAY」シートと「AE」シートを用意する．「SUMMARY」シートの帳票作成までの処理フローを以下に示す．「AE」シートも同様のフローとなる．

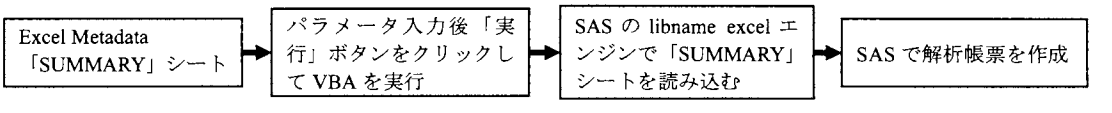


図 2.9 被験者背景要約表作成フロー

「SUMMARY」シート及び「AE」シートには，Item 列に各項目の名前が表示されており，Dataset 列で Dataset Metadata に記載されているデータセットを選択し，Variable 列で変数を選択する．同一ファイル内に図 1.2 に示す各 Metadata に関するシートが含まれているため，Variable 列で変数を選択する際に，Excel の VLOOKUP

関数等を用いて当該変数のラベルやコードリストなどの情報を自動的に表示し，入力作業を簡略化する機能を実装している．また，レコードフラグ変数の値や，解析帳票のタイトル等，テキスト入力が必要な項目については，Value 列に値を入力する（文字変数の場合は引用符で囲む）．入力完了後，「実行」ボタンをクリックする．

SUMMARY シート（被験者背景の要約表作成ツール）：
シート上部で投与群やレコード選択フラグ及び解析変数（入力必須）を指定する．年齢などの連続量に対しては，右下部で算出する要約統計量を選択する．

Item	Dataset	Variable	Label	Display Format	Codelist	Value	MVNAME
Treatment Group	ADSL	TRT01PN	Planned Treatment (N)	1.0	TRTN		TRT
Selection Criteria 1	ADSL	FASFL	Full Analysis Set Population Flag	$10.	NY	"Y"	SELECT1
Selection Criteria 2							SELECT2
Selection Criteria 3							SELECT3
Analysis variable 1	ADSL	AGE	Age	3.0			VAR1
Analysis variable 2	ADSL	SEX	Sex	$1.	SEX		VAR2
Analysis variable 3	ADSL	RACEN	Race (N)	1.0	RACEN		VAR3
Analysis variable 4							VAR4
Analysis variable 5							VAR5
Summary Statistics						N MEAN STDDEV MIN MEDIAN MAX	STAT
Title 1						Table 1.1	TITLE1
Title 2						Demographic and Baseline Characteristics	TITLE2
Title 3							TITLE3

実行

TRUE	☑ N	N
TRUE	☑ Mean	MEAN
TRUE	☑ SD	STDDEV
TRUE	☑ Min	MIN
FALSE	☐ Q1	
TRUE	☑ Median	MEDIAN
FALSE	☐ Q3	
TRUE	☑ Max	MAX

AE シート（有害事象の要約表作成ツール）：
投与群，MedDRA コーディング変数及び AE 発現フラグ（入力必須），レコード選択フラグ等を指定する．

Item	Dataset	Variable	Label	Display Format	Codelist	Value	MVNAME
Treatment Group	ADSL	TRT01AN	Actual Treatment (N)	1.0	TRTN		TRT
Selection Criteria 1	ADSL	SAFFL	Safety Population Flag	$10.	NY	"Y"	SELECT1
Selection Criteria 2							SELECT2
Selection Criteria 3							SELECT3
1st AE Flag	ADAE	AOCCFL	1st Occurrence of Any AE Flag	$1.	Y	"Y"	AEFLAG
1st SOC Flag	ADAE	AOCCSFL	1st Occurrence of SOC Flag	$1.	Y	"Y"	SOCFLAG
1st PT Flag	ADAE	AOCCPFL	1st Occurrence of Preferred Term Flag	$1.	Y	"Y"	PTFLAG
Other Record Flag 1							RFLAG1
Other Record Flag 2							RFLAG2
SOC CODE	ADAE	AEBDSYCD	Body System or Organ Class Code	8.0			SOCCODE
SOC NAME	ADAE	AEBODSYS	Body System or Organ Class	$200.			SOCNAME
PT CODE	ADAE	AEPTCD	Preferred Term Code	8.0			PTCODE
PT NAME	ADAE	AEDECOD	Dictionary-Derived Term	$200.			PTNAME
Title 1						Table 2.1	TITLE1
Title 2						Adverse Events	TITLE2
Title 3							TITLE3

実行

図 2.10 解析帳票作成シート

「SUMMAY」シート及び「AE」シートの実行結果を図 2.11 及び図 2.12 に示す．要約統計量や発現頻度等の計算結果とともに，指定した変数のラベル，コードリスト等が出力されていることが確認できる．

Table 1.1
Demographic and Baseline Characteristics

Variable	Statistics /Categories	Drug A	Drug B	Total
Age	N	145	133	278
	Mean	58.5	56.7	57.6
	SD	13.5	12.9	13.2
	Min	22	28	22
	Median	59.0	58.0	59.0
	Max	88	84	88
Sex	Female	44(30.3)	42(31.6)	86(30.9)
	Male	101(69.7)	91(68.4)	192(69.1)
Race (N)	ASIAN	145(100.0)	133(100.0)	278(100.0)

図 2.11 被験者背景の要約表

Table 2.1
Adverse Events

System Organ Class /Preferred Term	Drug A	Drug B
ALL	18 (12.0)	19 (13.7)
Gastrointestinal disorders	9 (6.0)	8 (5.8)
Constipation	3 (2.0)	1 (0.7)
Diarrhoea	0 (0.0)	2 (1.4)
Mallory-Weiss syndrome	0 (0.0)	2 (1.4)
Vomiting	2 (1.3)	0 (0.0)
Abdominal pain	1 (0.7)	0 (0.0)
Acute abdomen	1 (0.7)	0 (0.0)
Epigastric discomfort	1 (0.7)	0 (0.0)
Faeces hard	0 (0.0)	1 (0.7)
Gastric polyps	1 (0.7)	0 (0.0)
Gastric ulcer haemorrhage	1 (0.7)	0 (0.0)
Large intestine polyp	0 (0.0)	1 (0.7)
Toothache	0 (0.0)	1 (0.7)

図 2.12 有害事象の要約表

2.6 SAS プログラムによる OpenCDISC の実行

Jack Shostak ら（2012）は，SAS プログラムから OpenCDISC を実行して SDTM 及び ADaM データセットについてバリデーションを行う方法を提案している．本稿では，その方法をもとに，「Validation」シートにパラメータを入力し，VBA と SAS プログラムから OpenCDISC v1.5 を実行する方法を提案する．「Validation」シートに入力する項目を以下に示す．

項目	内容
Source data path	データセット（xpt ファイル）の格納フォルダのフルパス
OpenCDISC path	OpenCDISC のメインフォルダのフルパス
jar file name	jar ファイルのフルパス
data files	バリデーションを行うデータ（*.xpt で全てのデータセットを指定し，単一のファイルや LB*.xpt のように指定することも可能）
Config file	Config ファイルのフルパス
Define.XML	バリデーション時の Define.XML の使用の有無（Y またはブランクで，Y の場合は Define.xml を Data フォルダに格納）

表 2.2 「Validation」シートの入力項目

続いて，「Validation」シートのレイアウトを以下に示す.

Item	Value		MVNAME
Source data path	C:\Temp\SUGI_2014\Data	参照	_SOURCE
OpenCDISC path	C:\Temp\SUGI_2014\OpenCDISC\1.5\opencdisc-validator	参照	_OPENCDISCPATH
jar file name	C:\Temp\SUGI_2014\OpenCDISC\1.5\opencdisc-validator\lib\validator-cli-1.5.jar	参照	_JAR
data files	*.xpt		_FILES
Config file	C:\Temp\SUGI_2014\OpenCDISC\1.5\opencdisc-validator\config\config-adam-1.0.xml	参照	_CONFIG
Define-XML	Y		_DEFINE

実行

図 2.13 「Validation」シート（OpenCDISC 実行ツール）

「実行」ボタンクリック後，VBA から SAS が実行され，Output フォルダに OpenCDISC 実行結果のレポートが作成される.

OpenCDISC Validator Report

Configuration: C:\Temp\SUGI_2014\OpenCDISC\1.5\opencdisc-validator\config\config-adam-1.0.xml
Define.xml: Not provided
Generated: 2014-06-23T19:08:33
Engine Version: 1.5

Processed Sources

Domain	Label	Class	Source	Records	Errors	Warnings	Notices
GLOBAL	Global Metadata	–	–	–	0	0	0
ADAE	Basic Data Structure	Basic Data Structure	adae.xpt	55	3	1	2
ADEF	Basic Data Structure	Basic Data Structure	adef.xpt	852	0	0	1
ADSL	Subject-Level Analysis	Subject-Level Analysis	adsl.xpt	300	4	0	8
Total				1207	7	1	11

Unprocessed Sources

Domain	Label	Class	Reason	Errors	Warnings	Notices
Total				0	0	0
Grand Total				7	1	11

図 2.14 OpenCDISC 実行結果

以上のように，SAS，Excel Metadata 及び VBA を組み合わせることで，Define.xml，解析帳票の作成及びバリデーション等の ADaM に関連する作業を効率的に実施することが可能となる.

3 まとめ

本稿では，ADaM の Define.xml 作成時に必要な情報を組み込んだ Excel 形式の Metadata を作成し，さらに Excel VBA と SAS を組み合わせて Define.xml や解析帳票の作成ならびに OpenCDISC を用いたバリデーション作業を効率的に実施する方法を提案した.

PMDA から電子データ提出についての通知が発出されたこともあり，本邦においても新薬承認申請時の電子データ提出義務化への動きは加速し，製薬企業の負担は増大することが予想される．また，製薬企業が CRO に CDISC 標準に関する業務を委託する場合が多いと想定されるが，規制当局に臨床試験の電子データを提出することは製薬企業にとって非常に大きな責任を伴う業務であり，成果物の仕様の標準化や規制当局との協議ならびに成果物を受け取った際の自社内での受け入れ確認等を重厚に行う状

況も想定される．そのため，ADaM のみならず，CDISC に関連する臨床試験データや文書作成作業の効率化は製薬企業にとって重要な課題となる．その中で，本稿で紹介したような，実務担当者が日常的に用いる SAS や Excel 等の身近なツールの利便性を最大限に活用することは非常に有用であると考える．

連絡先

本稿で用いた Excel Metadata や SAS プログラム等に関する質問は下記に連絡されたい．
youhei.takanami@takeda.com

参考文献

- FDA Study Data Standards Resources
  http://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm
- FDA (2014). STUDY DATA TECHNICAL CONFORMANCE GUIDE (DRAFT)
  http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf
- PMDA 次世代審査・相談体制について（申請時電子データ提出）
  http://www.pmda.go.jp/operations/shonin/info/iyaku/jisedai.html
- PMDA（2014）薬食審査発第 0620 第 6 号通知「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」
  http://www.pmda.go.jp/operations/shonin/info/iyaku/jisedai/file/140620-tsuchi.pdf
- CDISC http://www.cdisc.org/
- Chris Holland, Jack Shostak (2012). Implementing CDISC Using SAS: An End-to-End Guide. Sas Inst
- 高浪 洋平（2013）. Simple Tool for Creating ADaM Define.xml for Statisticians in Pharmaceutical Companies Using SAS and HTML Application with Excel Metadata File. CDISC 2013 Japan Interchange
- 森岡 裕（2013）「ライブラリ参照と名前定義を利用して EXCEL ファイルへの柔軟な入出力を実現する方法と応用例の提案―解析結果のレポーティングからセルオートマトンまで―」SAS ユーザー総会 2013
- 高浪 洋平, 舟尾 暢男（2012）「統計解析ソフト『SAS』」工学社
- OpenCDISC Validator http://www.opencdisc.org/

大学における統計家育成のためのCDISC教育の実践

佐野　雅隆
東京理科大学工学部

Implementation of CDISC Education for Biostatistician Training Course in University

SANO Masataka
School of Engineering, Tokyo University of Science

要旨

大学における統計家教育に際して，CDISC (Clinical Data Interchange Standards Consortium)標準，特にADaM (Analysis Data Model)を理解する上で基本となる考え方は何か，標準化の目的・意義と関連して述べる．

キーワード：ADaM，教育，Official reaining course

はじめに

CDISC(Clinical Data Interchange Standards Consortium)標準の医薬品開発における重要性が増している．その中でも，Study Data Tabulation Model（SDTM）と Analysis Data Model（ADaM）は，PMDA に提出することから，その重要性が高まっていると考えられる．

産業界では，CDISC に関する教育研修の機会を設け，社員の能力開発に努めている．東京理科大学では医薬統計家の育成に取り組んでいるが，CDISC に関しては取り組み途中である．CDISC 標準に関して医薬統計家が具備すべき知識の全体像はまだ明らかではない．

研究室内での有志メンバーによる CDISC 教育(勉強会)を開始したので，とくに，医薬統計家に向けた ADaM の導入教育の実践について報告する．

産業界における教育内容の調査

産業界では，CDISC に関する教育研修の機会を設け，社員の能力開発に努めている．大学において教育を実施するに当たり，これらの事例を参考にするため，調査を実施した．浅見ら[1]は，統計解析担当者に対する CDISC の社内教育として，下記の教育プランを提案している．

1. 関連する SDTM/ADaM ドメイン横断の教育
2. SAS を用いたハンズオン
   (ア) サンプル ADaM データセットのレビュー
   (イ) ADaM に準拠した標準化 SAS 解析プログラムの実行
3. eCTD，申請パッケージを用いた説明

(ア) SAS XPT がどのように eCTD に添付されているか

(イ) FDA が実際にどのようにデータセットをレビューしているか?

さらに，SAS 社による CDISC 概論，SAS による CDISC SDTM/ADaM の実装では，

1. CDISC の概要
2. データモデルについて
3. ADaM メタデータと ADaM Define.xml
4. ADaM の実装
5. ADaM データのバリデーション

また，ある企業では，下記の取り組みをホームページ上で紹介している．

●CDISC 標準モデルと IG の環境

●各種検討会・実装化の準備

●教育資料の作成・社内への展開

●今後のトレーニング・教育研修対策

教育項目の選定と実施

浅見らの提案する教育項目は，グローバル申請の際に，併合解析を実施することを念頭に置いている．さらに，社内で ADaM のガイダンス教育を一通り実施した後の課題認識に基づいて立案されたものである．大学における統計家育成においては，ADaM のガイダンスを一通り実施することがまず必要であると考えた．一方で，ガイダンスの全体を詳細に理解することに比べて，概要を理解した上で標準化の目的や標準化を用いた上での臨床試験の質の担保・向上に向けた利活用への理解が重要であると考え，教育内容を選定することにした．

ADaM における中心的なデータ構造としては, ADSL (Subject level analysis dataset), BDS(Basic data structure), ADTTE(BDS for time-to-event analysis)，ADAE (ADaM data structure for adverse event analysis)が挙げられる．上記の目的から，まずは，ADSL，BDS に焦点を当てることにした．

教育を実施するにあたり，まずは，CDISC の公式トレーニングに参加した．その後，CDISC の文書として参考にしたものは，Analysis Data Model (ADaM) v2.1，Analysis Data Model (ADaM) Implementation Guide v1.0, Analysis Data Model (ADaM) Examples in Commonly Used Statistical Analysis Methods の 3 種類について，その内容の中から抜粋した．

たとえば，Analysis Data Model (ADaM) Implementation Guide v1.0 からは，下記の内容を抽出した．

2 Fundamentals of the ADaM Standard

2.1 Fundamental Principles

2.2 Traceability

2.3 The ADaM Data Structures

2.3.1 The ADaM Subject-Level Analysis Dataset ADSL

2.3.2 The ADaM Basic Data Structure (BDS)

3.1 ADSL Variables

3.2 ADaM Basic Data Structure (BDS) Variables

3.2.2 Treatment Variables for BDS Datasets

3.2.4 Analysis Parameter Variables for BDS Datasets

標準化の目的や標準化を用いた上での臨床試験の質の担保・向上に向けた利活用においては,

- 標準化と改善の基礎・PDCA, SDCA サイクル
- 質の定義
- 技術標準としての CDISC の位置づけ
- 管理指標

について, 触れることにした.

Examples in Commonly Used Statistical Analysis Methods には, 下記の手法が取り上げられており, 統計か教育においては実例をイメージしやすいと思われたので, 上記の基本的な知識の習得後に, データセットの作成, および解析の試行にとりかかることとした.

- Analysis of Covariance (ANCOVA)
- Analysis of Variance (ANOVA)
- Chi-squared
- Chi-squared, Corrected
- Cochran-Mantel-Haenszel
- Mantel Haenszel
- Fisher's Exact
- Kruskal-Wallis
- Log Rank
- McNemar
- Regression, Cox
- Regression, Linear
- Regression, Logistic
- Sign Test
- t-Test, 1-sided
- t-Test, 2-sided
- Wilcoxon (Mann-Whitney)

SDTM-ADaM pilot Study のデータを基にして, SAS を用いたハンズオンであるサンプル ADaM データセットのレビュー, ADaM に準拠した標準化 SAS 解析プログラムの実行を試行することが理想的であったが, SDTM-ADaM の Pilot データを入手することは困難であったため, データの作成も踏まえて検討することにした.

以上の検討を基にして, 参加者を募集した. 研究室全体に向けて勉強会の案内を通知した, とこ単位を取得できるコースではないにも関わらず, 修士の学生 7 名の参加者が集まった. 一方で, 社会人学生の方や, 研究室を卒業して産業界で活躍している方の中にも, 興味を持った方もいたものの, 教育実施の時間帯の調整が難しく, まずは学内のみで実施することにした.

参加者には, 各種ドキュメントを配布したが, 英語で書かれたドキュメントであるため, 必ずしも理解がスムーズに進むとはいえなかった. 現在, 導入教育が終了した段階であり, 今後 example を基に, データセットの作成, および解析の試行に移行する予定である. 大学・産業界からの外部講師の招聘を含めて計画したい.

今後の課題

教育項目の妥当性について，産業界からの意見を反映することが必要であると考えている．現状の教育について，計画を実行することとその評価が必要である．さらに，ICH E9 の Trial Statistician の役割の明確化と CDISC 標準に関する教育項目との対応付けが挙げられる．東京大学における教育内容の把握も必要である．

参考文献

[1] 浅見由美子ら，"SAS を用いた医薬品開発の統計解析担当者に対する CDISC の社内教育"，SAS ユーザー総会 2013 論文集，423-438.

[2] Analysis Data Model (ADaM) v2.1, CDISC,2009

[3] Analysis Data Model (ADaM) Implementation Guide v1.0, CDISC,2009

[4] Analysis Data Model (ADaM) Examples in Commonly Used Statistical Analysis Methods, CDISC,2009

[5] Chris Holland and Jack Shostak. Implementing CDISC Using SAS, 2012

[6] 東京大学 SPH シラバス，http://www.m.u-tokyo.ac.jp/education/4_sph_2014.pdf

産官学でのCDISC標準利用に向けた取り組み：実務担当者のために

承認申請のためのCDISC実装とメタデータ作成

氏名

浅見由美子（第一三共）、奥田恭行（第一三共）、Tony Chang（Amgen Inc.）

Efforts toward the utilization of CDISC standards in industry, government, and academia

Implementation of CDISC Standards and the Metadata for Regulatory Submission

Name

Yumiko Asami[1)], Yasuyuki Okuda[1)], Tony Chang, Amgen Inc. [2)]

1) Daiichi Sankyo Co., Ltd., 2) Amgen Inc.

要旨

Clinical Data Interchange Standards Consortium（CDISC）とは、医薬品に関する臨床・非臨床データの標準を推進する非営利団体である。2016年（予定）より、PMDA承認申請の際にCDISCのデータ標準であるSDTM（申請臨床試験データモデル：Study Data Tabulation Model）/ ADaM（統計解析データモデル：Analysis Data Model）に基づく電子データセットの提出が義務化されることが公表されている。CDISC 標準に基づくデータセット（以下、CDISCデータセット）の作成には、CDISCガイダンスに基づいたメタデータ（データセットの仕様）が必要となる。メタデータにはCDISCガイダンスをそのまま適用すれば作成できる部分もあるが、承認申請のスポンサー（以下、スポンサー）が作成しなければならない部分も多く含まれる。承認申請用の併合解析や、SDTM/ADaM 作成用の SAS プログラムを効率よく構築するためには、臨床試験単位ではなく、プロダクトもしくはスポンサー内で共通で使用できるメタデータを作成することが重要となってくる。本稿では、グローバル承認申請におけるCDISC実装の経験をもとに、プロダクト内で適切に効率よくメタデータを構築する方法を提案する。あわせて、メタデータ作成における課題についてもまとめる。

キーワード：CDISC、SDTM、ADaM、メタデータ、SAS、承認申請

1 背景

1.1 CDISC

Clinical Data Interchange Standards Consortium（以下、CDISC）とは、医薬品に関する臨床・非臨床データの標準を推進する非営利団体である。[1)] CDISCが作成した標準（以下、CDISC標準）には、以下のようなモデルが含まれる。

- Study Data Tabulation Model (SDTM)：医薬品の規制当局の申請臨床試験データ形式
- Analysis Data Model (ADaM)：医薬品の規制当局の承認申請統計解析データ形式
- その他、CDASH、SEND、PR等

代表的な例として、Electric Data Capture (EDC)等で収集された臨床試験データからSDTMを経て、SAS等を用いてADaMデータセットが作成される。そしてADaMデータセットを基に、SAS等を用いて統計解析結果が生成される。図1に、SASとSDTM/ADaMを用いた統計解析プロセス例（模式図）を示す。

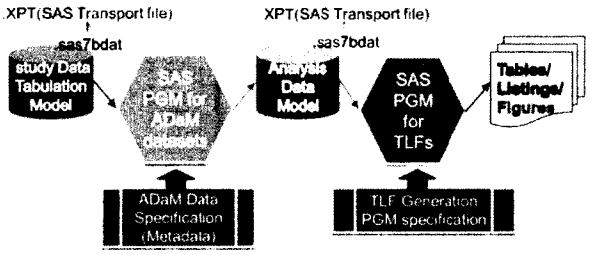


図1：SASとSDTM/ADaMを用いた統計解析プロセス例（模式図）

1.2　国際共同開発

「国際共同治験に関する基本的考え方」[2]の発行により、日本が参画する国際共同試験の数が急激に増加し、複数のスポンサー（製薬企業またはアカデミア）や複数の地域（日本、アメリカ、ヨーロッパ等）が開発計画に関与し、同じ時期に世界の各規制当局に対する承認申請を実施する場合もある。その結果、日本の製薬企業、アカデミアおよびCRO等の組織においても、FDA承認申請のために、日本で実施した臨床試験のCDISCデータセットを作成したり、アメリカで実施された臨床試験のCDISCデータセットを用いて、日本における承認申請のための解析を実施したりする場面も増加してきた。

1.3　医薬品承認申請における電子データ提出に関する各規制当局の対応

以下に各規制当局における医薬品承認申請における電子データ提出に関する対応を簡単に述べる。

FDAは1980年代からSASデータセット（その後、SAS Transport v5 format）を移送形式とした電子データ提出をスポンサーに要求し、さらに2004年からは、データセットの構造、変数の属性等の標準としてCDISC標準を推奨してきた。そして、2014年に「Draft Guidance for Industry Providing Regulatory Submissions in Electronic Format Standardized Study Data」ならびに「Draft Study Data Technical Conformance Guide」を発行した。それにより、FDAの医薬品承認申請におけるCDISC標準に基づく電子データ提出の強化が予想される。[3]

一方、PMDA（医薬品医療機器総合機構）は臨床試験の電子データ提出を従来スポンサーに要求していなかったが、最近CDISC SDTM/ ADaM標準に基づく電子データ提出を2016年（予定）より義務化することを公表した。これにより、日本の製薬企業、アカデミアおよびCRO等の組織においても、PMDAの承認申請のためにCDISCデータセットを作成し、提出することが必要となる。[4]

2　CDISCデータセットのメタデータ作成における問題点

2.1　CDISCデータセット作成におけるメタデータ

メタデータ（metadata）とは、一般的には「データに関するデータまたは情報」を意味する。CDISC標準の実装においては、データセット、変数等に関する定義、仕様書にあたり、SAS等を使用してSDTM/ADaMデータセットを作成（変換）する際のベースとなる。メタデータは、SDTM Implementation Guide (以下、SDTM IG)、ADaM Implementation Guide（以下、ADaM IG）等のCDISC標準に関するガイド類[1]に従う必要がある。承認申請時には「Define file」に含めて、規制当局へ提出される。なお、本稿においては、メタデータはSDTM/ADaMデータ作成（変換）の元になる仕様書を意味し、Define fileの意味を含めないこととする。

図 2 に SAS を用いて作成した場合の SDTM AE ドメインの一部（見本）を示す。メタデータの検討の際には、データセット（ドメイン）レベル、および、変数レベルで検討する必要がある。

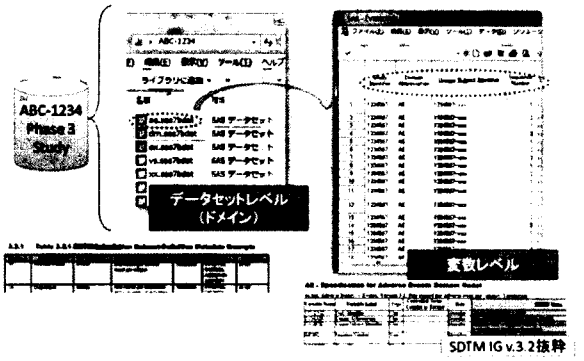


図 2：SAS を用いて作成した場合の SDTM AE ドメインの一部（見本）

2.2　メタデータ作成における問題点

日本製薬工業協会（以下、製薬協）が 2013 年に実施したアンケート[5]によると、製薬協加盟企業の約半数が SDTM 実装の経験があるが、そのうち、その約 3/4 は SDTM 標準に適合しているとは言えない（いわゆる「SDTM Like」）であるとのことであった。それらの企業では、SDTM データセット作成の基となるメタデータが SDTM IG に適合していない可能性がある。

EDC 等で収集された臨床試験データ（以下、CRF データ）を基に SDTM データセットを作成するためには、SAS 等を用いたデータ加工（複数データセットのマージや転置等の構造変換、変数の属性変更、単数もしくは複数の変数の値を代入した四則演算や関数等を用いた変換など）が必要となる場合がある。CRF データの構造によっては、データ加工の仕様が複雑になることもあるため、SDTM メタデータ作成のためには SAS 等によるデータ加工のプログラミングの知識も必要となる。SDTM 実装の経験が少ない組織では、SDTM メタデータ作成における SAS プログラミング知識の必要性が認識されていない場合もあり、SDTM 作成を担うファンクションに SAS プログラミングの知識のある人材が不足している可能性もある。

また、SDTM IG や ADaM IG には画一的ではなく、複数のオプションを取りうる事項が含まれている。そのことが、メタデータ作成を困難にしている可能性がある。

3　医薬品の承認申請における統計解析とデータ標準化

図 3 に SAS と ADaM を用いた承認申請用解析のプロセス例（模式図）を示す。使用する臨床試験数で出力する解析帳票を分類すると以下ようになる。

1)　単一の臨床試験のデータを用いる解析帳票（主に総括報告書用）

2)　複数の臨床試験のデータを用いる解析帳票（主に承認申請資料用）

複数の臨床試験データを用いる解析帳票には、例として併合解析や複数試験の「横並び」の解析（臨床試験、投与群を横に並べたような解析）がある。

臨床試験間で解析データの仕様が統一されていない場合、1)において解析プログラムを再利用することが難しくなり、臨床試験毎に解析帳票作成プログラムを作り直すことが必要となる。また 2)において解析を実施する前に、承認申請資料に含まれる複数の臨床試験の解析データの仕様をそろえるための変換が必要となる。その変換作業は煩雑になることも多く、承認申請タイムラインに影響を与える場合もある。

そこで解析データ仕様の標準化が重要となってくる。CDISC 標準は規制要件としてだけでなく、上記のような問題を回避するためにもメリットがあると考えられる。更に、レビューアー（社内の関係者、規制当局

審査官等）がデータをレビューしやすくなるというメリットもあると考えられる。

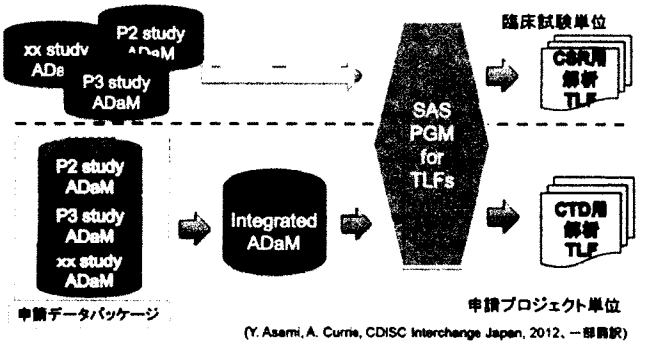


図 3：SAS と ADaM を用いた承認申請用解析のプロセス例（模式図）

4　メタデータの作成手順と管理項目

4.1　メタデータの記述パターン

SDTM と ADaM では、より自由度の低い SDTM メタデータの IG への適合性がより重要であると考えられるため、SDTM メタデータを中心に検討を進める。実際の SDTM 実装の経験を元に、表 1 に SDTM メタデータにおける記述のパターンの分類を示す。本稿においては 4 つの分類とした。CRF データの構造に依存し 1)に該当する割合が高くなり、メタデータの作成、及び、その後のプログラミング作業の負担が大きくなる。また、SDTM の経験が少ない組織の場合は、2)、3)に該当するメタデータ作成、および、その SDTM IG への適合性の確認に時間が掛かる可能性がある。

表 1：SDTM メタデータの記述パターンと代表的なデータ加工例

メタデータの記述パターン	代表的なデータ加工例 （データ構造変換、属性の変更、派生等）
0) CRF データの Copy	－
1) SDTM IG で決定されている属性と CRF データのギャップの穴埋め	Dataset Name、Structure、Key variable の変更 Variable Name, label, Type の変更 Controlled Terminology の変更、等
2) 一般的に CRF データとして収集されておらず、他のマテリアルの記述に依存するもの	Trial Design Terminology の作成 Baseline Flag の作成 Population Flag の作成, Extensible Terminology の追加 Sponsor Specified Terminology の追加、等.
3) その他、一般的にスポンサーが決定する属性	Custom Domain Name, Structure の設定（Class の選択） Variable Length の調整 Supplemental Qualifier の作成、等

*注：2)と 3)は重複もありうる。

4.2　医薬品承認申請を考慮したメタデータ作成手順

表 1 のメタデータの記述パターンを基に、医薬品の承認申請時の工数軽減も考慮した、メタデータ作成の負担軽減策を表 2 にまとめる。

表 2：SDTM メタデータの記述パターンと負担軽減策

メタデータの記述パターン	負担軽減策
0) CRF データの Copy	—
1) SDTM IG で決定されている属性と CRF データのギャップの穴埋め	・CDASH の活用、プロジェクトまたは組織内の CRF データ標準化、英語化（日本ローカル試験の場合）
2) 一般的に CRF データとして収集されていない項目	・SDTM を考慮したプロトコル標準化、、マシンリーダブル形式の採用、英語化（日本ローカル試験の場合） ・SDTM を考慮した統計解析早期策定(ベースラインの定義、解析集団の定義、等）等 ・図 4 及び表 3 参照
3) 一般的にスポンサーが決定する属性	・図 4 及び表 3 参照

表 2 における 2)及び 3)、及び、ADaM メタデータは、スポンサーで決定できる属性を含むため、承認申請プロジェクト内の臨床試験間でばらついてしまう場合がある。そのことが承認申請用の解析実施時の工数増大につながる可能性がある。また、臨床試験ごとにメタデータを作成する場合、本来同じ派生ルールの場合でも、記述方法が異なってしまう可能性もある。

そこで、図 4 のイメージにあるような、承認申請プロジェクト内の臨床試験のメタデータを横断的に管理できるようなシートを用いることとした。一般的に、承認申請プロジェクトに含まれる臨床試験は実施時期が異なるため、新規の臨床試験時はカラムを追加する形式とした。また、そのシートを用いたメタデータの管理項目と作成手順を表 3 に示す。

	A	B	C	D	E	F	G	H	T	J	K
1	Dataset	Variable Name	Variable Label	Variable Type	Display Format	Codelist/ Controlled Terms	General Source/Der ivation	Core	[redacted]		
2	ADxx	aaa	zzzz	Char	$zz	zzzz	zzzz	Req	x	x	If xxx=xxx then aaa="X"
3	ADxx	bbb	zzzz	Char	$zz	zzzz	zzzz	Cond	NA	x	NA
4	ADxx	ccc	zzzz	Char	$zz	zzzz	zzzz	Req	x	x	x
5	ADxx	ddd	zzzz	Num	zz	zzzz	zzzz	Cond	NA	NA	x

図 4：承認申請パッケージ内のメタデータ管理シート（イメージ）

表 3：承認申請パッケージ内のメタデータ管理項目と作成手順

順序	レベル	管理項目
1	データセット（ドメイン）レベル	Dataset Name, Structure, Key Variable(s) *必要な場合はカスタムドメインの設計
2	変数レベル	Variable Name, Label, Type, Display format, Derivation Rule *特に CRF データの構造が臨床試験間で異なる場合は変数派生ルール(Derivation Rule)の確認が重要
3	ターミノロジー、辞書	Controlled Terminology (Extensible or not) Sponsor Specified (Trial design, etc.) *臨床試験間で別の意味のデータに、同じ Terminology を当てはめないように注意

5　実際の承認申請における適用

図 4 のシートと表 3 に示す方法を用いて、実際のグローバル承認申請プロジェクトにおけるメタデータを作成、管理した。その結果をまとめると以下のとおりである。

- 既に存在する変数はそのまま使用することができた。また、既に存在する変数を誤って別の名前または属性で新規に作成することが防げた。
- 臨床試験間で同じ意味の場合、Terminlogy を共通して使用することができた。一方、別の意味の場合、誤って臨床試験間で同じ Terminology にならないように、排他的に作成できた。
- 変数派生ルールの記述を再利用することができた。また、それに基づく SAS コードを再利用することができた。
- 異なる会社、地域間でメタデータを見える化、共有化することができた。

6　まとめ

SDTM/ADaM メタデータを IG に適合する形で、効率よく作成するためには CDASH およびプロトコル標準を用いて、End-to-End の標準化をすることが望ましい。しかし、組織全体でそのようなプロセスに移行するにはある程度の時間が掛かる。また、医薬品の承認申請準備の効率化も考慮すると CDASH の適用のみでは解決できない問題点もある。そこで、本稿では実際の SDTM メタデータ作成の経験を基に、SDTM メタデータの記述パターンを分類することにより、CDASH の適用で負担を軽減できる部分と、その他の方策も必要な部分を明確にした。CDASH 以外のその他の方策の一つとして、医薬品承認申請を考慮したメタデータ作成手順と管理項目を設定した。そして、実際のグローバル承認申請でその方法を用いた結果をまとめた。

PMDA 承認申請における 2016 年（予定）からの CDISC データセット提出の実現のためには、End-to-End で標準化されたプロセスへの移行準備と、それが定着するまでの過渡期の対応を同時に進めていく必要があると考える。

＜参考＞

1) CDISC Website http://www.cdisc.org/

2) 薬食審査発第 0928010 号. 国際共同治験に関する基本的考え方；2007

3) FDA Study Data Standards Resources
http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm

4) 医薬品医療機器総合機構　次世代審査・相談体制について（申請時電子データ提出）
http://www.pmda.go.jp/operations/shonin/info/iyaku/jisedai.html

5) 製薬工業協会発表資料、CDISC Interchange Japan ；2013

PMDAにおける次世代審査・相談体制とCDISCの利用について

安藤友紀
独立行政法人医薬品医療機器総合機構　次世代審査等推進室

Advanced New Drug Review and Utilization of CDISC in PMDA

Yuki Ando
Advanced Review with Electronic Data Promotion Group, Pharmaceuticals and Medical Devices Agency

要旨

近年の医薬品開発においては、開発の意思決定におけるデータに基づく定量的な情報の積極的な利用が進められている。本邦では、健康医療戦略（平成25年6月14日内閣官房長官、厚生労働大臣・関係大臣申合せ）において、医薬品医療機器総合機構（PMDA）における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が求められている。現在、PMDAでは臨床試験データの電子的提出を受けることを視野に、医薬品承認申請時に添付される臨床試験データ利用のための具体的検討を行っており、その基本的な方針について通知「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」（平成26年6月20日薬食審査発0620第6号）が発出されたところである。医薬品の承認審査への臨床試験データの利用、及び将来的な蓄積データの横断的な検討には、データ標準の利用が不可欠であり、PMDAでは、データ標準としてCDISCによる標準を採用する予定である。国際的に広く使用されている標準の採用により、PMDA及び申請者の両者において、国際連携を視野に入れたより適切かつ最先端の解析や評価が可能となると考えられる。本発表では、PMDAにおける次世代・審査相談体制の概要、準備状況と、実施中のパイロットを含むCDISCの利用に関する取り組みについて紹介する。

キーワード：承認申請、臨床試験データ、データ標準

次世代審査・相談体制

近年の医薬品開発においては、開発の意思決定におけるデータに基づく定量的な情報の積極的な利用が進められている。開発中に得られる多くの情報を数理モデルにより統合し効果等の予測に用いるModeling and Simulationの利用はその代表的なものである。一方、医薬品を開発し承認申請を行う者のみならず、審査当局においても臨床試験データ及び各種手法の理解、及び積極的な活用を進めることは重要であり、健康医療戦略（平成25年6月14日内閣官房長官、厚生労働大臣・関係大臣申合せ）においては、医薬品医療機器総合機構（PMDA）における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が求められている。PMDAでは、平成25年9月1日の「次世代審査・相談体制準備室」の設置、平成26年4月1日の「次世代審査等推進室」への改組を経て、現在、承認申請時に提出される臨床試験データを一層活用した承認審査や相談の実施について具体的な検討が進められているところであり、その基本的な方針に関する通知「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」（平成 26 年 6 月 20 日薬食審査発 0620 第 6 号）が発出されたところである。

承認申請時に提出される臨床試験データは、まずは各品目の審査の効率化、高度化のために利用される。PMDA のこれまでの審査においては、PMDA 自身が臨床試験の被験者レベルの電子データを持っていなかったことから、有効性及び安全性の検討において申請資料として提出されていない解析の結果を確認する必要が生じた場合、全ての解析を申請者に依頼し実施してもらう必要があった。しかしながら、被験者レベルの電子データが入手可能となることにより、一定の解析については PMDA 内で実施する等、申請者の負担が軽減される側面があると考えられる。また、ソフトウエアの利用による試験結果の視覚化や、試験結果から個別被験者のデータへのアクセスが容易になる等、臨床試験結果のより詳細な検討が可能となることも期待される。一方、承認申請時に提出される複数品目の臨床試験データが蓄積されることにより、例えば特定の疾患の治療薬に関する品目横断的な検討に有用な情報が得られる可能性もある。このような検討に基づくガイドラインの発出等、蓄積されたデータが将来的な医薬品開発の効率化につながることも期待される。

PMDA における CDISC の利用について

承認申請時に PMDA に対して臨床試験の電子データが提出される場合に、各臨床試験のデータが一定の標準に従って構成されているか否かは非常に重要となる。審査員は審査や解析に先立ち、個々の臨床試験のデータの内容を理解する必要があるが、データが特定の標準に従っている場合には理解が容易になると同時に、標準の使用を想定して作られたソフトウエアによりデータの視覚化等を容易に行うことができる。また、将来的な蓄積データの横断的な検討には、各臨床試験のデータが特定のデータ標準に準拠していることが不可欠となる。国際的なデータ標準への準拠は、医薬品の審査のみならず開発段階の国際的な連携等においても重要であると考えられる。

PMDA では臨床試験データのデータ標準として CDISC（Clinical Data Interchange Standards Consortium）による標準を採用することとし、SDTM（Study Data Tabulation Model）に従う臨床試験データ、ADaM（Analysis Data Model）に従う解析用データ、及びそれらの定義ファイル（Define.xml）の提出を求める予定である。PMDA における CDISC に準拠した臨床試験データの入手は初めてのことであり、各審査員の CDISC への理解を促進する必要がある。また、申請者に関しても、データ標準として CDISC が採用され、それに従ってデータがまとめられている例はまだそれほど多くないようである。PMDA においては平成 25 年後半より、実際に製薬企業から CDISC 標準に準拠してまとめられた臨床試験データを提出してもらい、PMDA 内で一定の解析が可能であることを確認するパイロットを実施しており、パイロットの実施を踏まえて各社の CDISC 準拠データ間に多少の違いがあることを認識している。このような違いが出てくる原因としては、様々な臨床試験に適用可能な標準そのものの特徴もあるが、標準への理解度や経験の違いも考えられる。今後、本邦において CDISC の利用を推進していくにあたっては、標準自体の理解や情報の共有を進めることが重要であると考える。

今後の展開

PMDA では現在、「基本的考え方」に続き、実務的な詳細に関する通知に含めるべき項目について整理しており、今後も関係者との意見交換、パイロットの実施等を踏まえて検討を進めていく予定である。特に CDISC 標準に準拠した臨床試験データに関しては、これまでに準拠したデータを作成したことのある経験者の意見も踏まえて、利用にあたっての注意や、SDTM、ADaM それぞれのモデルに関する作成時の留意点等

のより詳細な内容を提供することが重要であると考えており、実際に製薬企業及びCROの実務担当者の協力も得て、技術的な準拠ガイドに記載すべき内容も検討中である。本邦において承認申請時の臨床試験電子データ提出が予定されていること、CDISCによる各種標準については今後のバージョンアップも想定されることから、CDISCの利用に関してはガイド等の発出に加えて、利用者内の頻繁な情報共有も重要であると考えられる。今後も、臨床試験におけるデータの収集、整理や解析データセットを用いた結果の解析、承認申請の準備等、申請までの各段階でCDISC標準データに関わる実務担当者や、承認審査にデータを用いるPMDAも含め、各自の経験を踏まえたタイムリーな情報共有により、本邦におけるCDISCの普及が速やかに進められることが期待される。

参考文献

厚生労働省医薬食品局審査管理課長、「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」、平成26年6月20日薬食審査発0620第6号

社内標準策定でのCDISCの利用

坂上　拓
株式会社 中外臨床研究センター バイトメトリクス部　データサイエンスグループ 1T

Standardization of statistical analysis process based on CDISC standard

Taku Sakaue
Biometrics Dept. Data Science Group 1T, Chugai Clinical Research Center., LTD

1. 要旨

標準を導入することの最大のメリットは，業務の効率化，アウトプットの品質の向上，また関係者間で同じ言葉も用いることによるコミュニケーションの向上などが考えられる．

承認申請時の CDISC 標準に準拠した臨床電子データの提出義務化を控え，我々が CDISC 標準を導入する際，承認申請対応という目的と共に，如何に標準を導入することで得られる恩恵を享受するかが，導入の大きなポイントとなる．

本発表では，当社で CDISC 標準の ADaM を導入と共に検討された業務プロセスを元に，我々の統計解析業務に ADaM を導入することで想定される効果（効率面，品質面，コミュニケーション）と，その標準の維持・管理について紹介する．

キーワード：CDISC 標準, ADaM, 統計解析業務プロセス

2. はじめに

CDISC 標準の一つである ADaM (Analysis Result Model)は，臨床研究データの統計解析と，それに続く統計的な検討を行うためのデータモデルである．この ADaM は，CDISC 標準の臨床研究データの申請用データモデルである SDTM (Study Data Tabulation Model)のようにハードな標準の枠組みが規定されておらず，自由度が高いデータモデルであることが知られており，作成されるデータセットの数，そのデータ構造，作成される変数は，解析要件（仕様，解析帳票のレイアウト）によって多様に変化する．

3. 解析要件と ADaM / データ処理の関係

臨床検査値のシフトテーブルを例に，解析要件によって ADaM や，解析プログラムがどのように変わるかを説明する．

臨床検査のような，被験者に対し複数の測定項目・測定時点を有するデータを解析する際，BDS (Basic Data Structure)と呼ばれる構造を持つ ADaM を用いる．

表 1. 臨床検査値のシフトテーブル(1)

			Baseline					
			Group A N=xx			Group B N=xx		
Parameters	Visit	Shift	Low	Normal	High	Low	Normal	High
NEUTROPHILS	WEEK2	Low	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)
		Normal	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)
		High	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)
	WEEK3	Low	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)
		Normal	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)
		High	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)

表 1 のシフトテーブルを実現するには，以下のような変数を有する ADLB を作成する．

ADLB

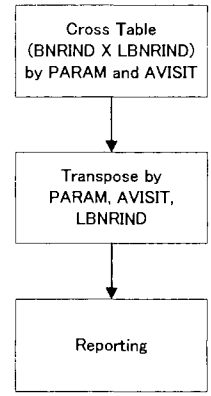
USUBJID	TRTA	PRAMCD	PARAM	ABLFL	AVISIT	BNRIND	LBNRIND
TEST-01	Group A	NEUT	NEUTROPHILS	Y	WEEK1	NORMAL	NORMAL
TEST-01	Group A	NEUT	NEUTROPHILS		WEEK2	NORMAL	LOW
TEST-01	Group A	NEUT	NEUTROPHILS		WEEK3	NORMAL	NORMAL
TEST-02	Group A	NEUT	NEUTROPHILS	Y	WEEK1	NORMAL	NORMAL
TEST-02	Group A	NEUT	NEUTROPHILS		WEEK2	NORMAL	LOW
TEST-02	Group A	NEUT	NEUTROPHILS		WEEK3	NORMAL	NORMAL
TEST-02	Group A	NEUT	NEUTROPHILS		WEEK4	NORMAL	HIGH

この ADLB の変数を解析帳票に annotate した結果と，解析帳票を作成するために必要な，データ処理の概要を以下に示す．

ADLB annotate 結果

			Baseline		
PARAM			BNRIND		
Parameters	Visit	Shift	Low	Normal	High
NEUTROPHILS	WEEK2	Low	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)
		Normal	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)
	AVISIT	High	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)
			LBNRIND		
	WEEK3	Low	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)
		Normal	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)
		High	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)
	WEEK4	Low	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)
		Normal	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)
		High	XXX (xxx.x)	XXX (xxx.x)	XXX (xxx.x)

処理概要



次に解析要件を変更し，以下のような臨床検査値のシフトテーブルを作成する場合を検討してみる．

表 2. 臨床検査値のシフトテーブル(2)

Parameters	Shift	Group A N=xx	Group B N=xx
NEUTROPHILS	Low/Normal to High	XXX (xxx.x)	XXX (xxx.x)
	High/Normal to Low	XXX (xxx.x)	XXX (xxx.x)

表 2 のシフトテーブルは，ベースラインの測定値から，ポストベースラインに発生した最悪値への変動（同一被験者が High / Low の両方向に変動した場合は，それぞれにカウント）を，特に関心のある変動に着目して集計したものになる．本帳票を作成するために，以下のような変数を有する ADLB を作成する．

ADLB

USUBJID	TRTA	PARAM	ABLFL	AVISIT	AVAL	BASE	CHG	BNRIND	LBNRIND	DTYPE	SHIFT1
TEST-01	Group A	NEUTROPHILS	Y	WEEK1	4.2	4.2	0	NORMAL	NORMAL		
TEST-01	Group A	NEUTROPHILS		WEEK2	1.7	4.2	-2.5	NORMAL	LOW		
TEST-01	Group A	NEUTROPHILS		WEEK3	2.3	4.2	-1.9	NORMAL	NORMAL		
TEST-01	Group A	NEUTROPHILS		POST-BASELINE MIMINUM	1.7	4.2	-2.5	NORMAL	LOW	MINIMUM	High/Nomal to Low
TEST-01	Group A	NEUTROPHILS		POST-BASELINE MAXIMUM	2.3	4.2	-1.9	NORMAL	NORMAL	MAXIMUM	
TEST-02	Group A	NEUTROPHILS	Y	WEEK1	7.5	7.5	0	NORMAL	NORMAL		
TEST-02	Group A	NEUTROPHILS		WEEK2	8.2	7.5	0.7	NORMAL	HIGH		
TEST-02	Group A	NEUTROPHILS		WEEK3	9.1	7.5	1.6	NORMAL	HIGH		
TEST-02	Group A	NEUTROPHILS		WEEK4	1.7	7.5	-5.8	NORMAL	LOW		
TEST-02	Group A	NEUTROPHILS		POST-BASELINE MIMINUM	1.7	7.5	-5.8	NORMAL	LOW	MINIMUM	High/Normal to Low
TEST-02	Group A	NEUTROPHILS		POST-BASELINE MAXIMUM	9.1	7.5	1.6	NORMAL	HIGH	MAXIMUM	Low/Normal to High

この ADLB の変数を解析帳票に annotate した結果と，解析帳票を作成するために必要な，データ処理の概要を以下に示す．

ADLB annotate 結果

PARAM

Parameters	Shift SHIFT1	Group A N=xx
NEUTROPHILS	Low/Normal to High	XXX (xxx.x)
	High/Normal to Low	XXX (xxx.x)
LYMPHOCYTES	Low/Normal to High	XXX (xxx.x)
	High/Normal to Low	XXX (xxx.x)

処理概要

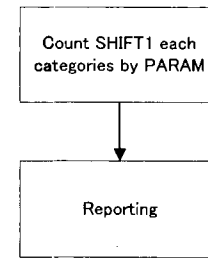


表 2 は，世間一般に「縦持ちデータ」と呼ばれる BDS 構造から作成し易いレイアウトで，データ処理の処理概要で減少した 1 ステップ以上に，解析帳票を実現するためのデータ処理が単純化できることは，表 1 のレイアウトの解析帳票を一度でも作成したことがあるプログラマなら容易に想像できる．

また，表 2 のレイアウトは，尿検査や医師所見といった定性値項目の頻度集計と同様のルーティンで処理することができ，解析プログラムの再利用も可能にする．

提示した ADLB やデータの処理概要はあくまで一例に過ぎないが，ADaM のデータ構造や変数の特性に応じて解析要件を工夫することにより，データ処理が単純化され，効率的に解析業務を遂行できるようになる．

4. 標準の活用

本章では，弊社で ADaM 導入時に検討された業務プロセス（図 1）を紹介すると共に，そのプロセスで得られる効果について紹介する．

この業務プロセスは，第 3 章のシフトテーブルの例で説明した，「解析要件によって作成されるデータセットの数，そのデータ構造，作成される変数が変わる」という考えの元，解析要件を CDISC に適した形で固定した上で，構築した業務プロセスになる．

まだ着手が始まったばかりで，これから示す業務プロセスは，構想段階の域を出ないことをご留意いただきたい．

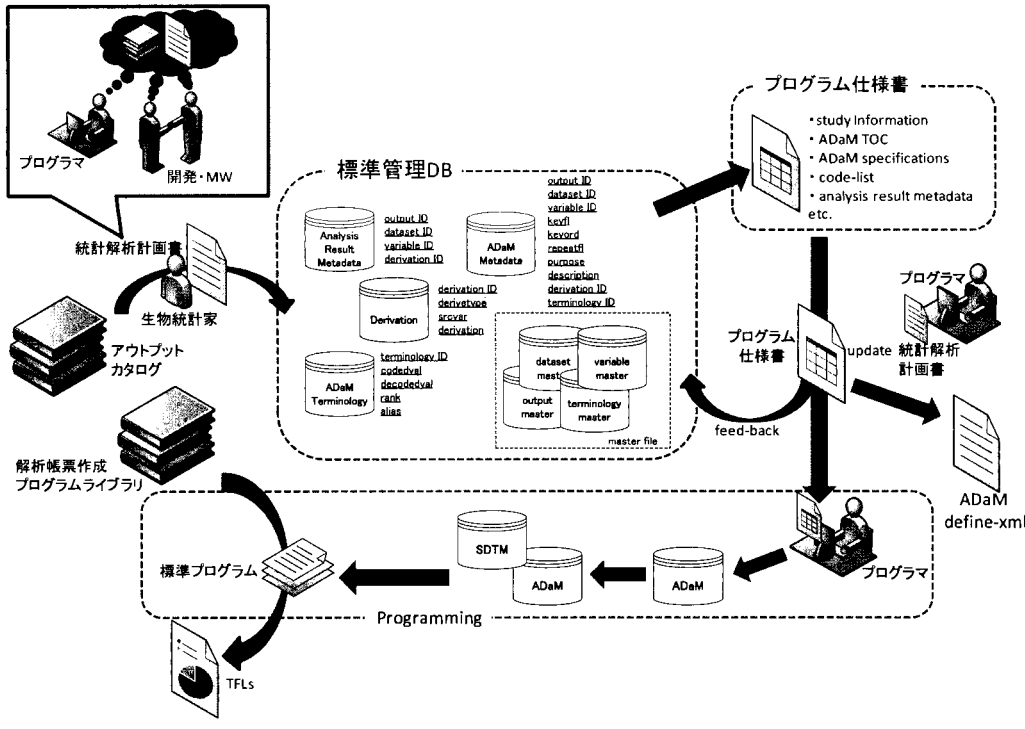


図 1. ADaM を導入した統計解析業務プロセス

4.1 アウトプットカタログ

アウトプットカタログは，どの試験でも概ね作成される解析帳票のレイアウト集で，レイアウト名とそのレイアウトが対になるように管理されていることが望ましい．生物統計家は，このアウトプットカタログから，必要な解析帳票を選択し，統計解析計画書を作成する（レイアウト集にないものは別途計画する）．

これにより，モックアップを作成する手間が省け，製品横断的に同じレイアウトの解析帳票を作成することができるようになる。また，アウトプットカタログを，臨床開発担当者やメディカルライター，プログラマと共有することで，出力イメージや，解析帳票に表示されている数値の意味を共有しやすくなると共に，レイアウト名を使った会話が成り立つようになる．

4.2 標準管理 DB

標準管理 DB は，アウトプットカタログのある解析帳票を実現するために必要な，ADaM のデータセットとそのメタデータ（変数属性や導出ルール，格納値のターミノロジー等），解析帳票のプログラム仕様を管理したものになる．当然，解析する疾患領域や，プロトコルによって解析要件は異なるため，全てを標準として管理するは不可能で，管理するものは必要最小限に留めておく（解析要件のバリエーションを把握できている場合は，バリエーション分管理した方が，より統制が効く）．この標準管理 DB は，標準管理者と標準を検討するチームを用意し，定期的に標準を見直すようなプロセスを構築して運用する．

また，この標準管理 DB と共に，解析要件に応じて標準から変更したもの，新たに追加したものを管理しておくと，定期的に標準を見直す際に参考情報となる．

4.3 プログラム仕様書と define-xml

プログラム仕様書は，統計解析計画書で計画されたレイアウト名を元に，標準管理 DB から自動的に生成する．これにより，社内で作成されるデータセットの仕様は統一されると共に，プログラム仕様を効率的に作成できることが期待される．

標準管理 DB から自動的に生成されたままの状態では，統計解析計画書の要件を満たさないため，統計解析計画書を元にプログラム仕様書を更新することになる．ここで更新された情報を，標準と比較し，その差分（標準からの変更点・新規に追加した仕様）を把握しておくことで，プログラム仕様書のレビューするポイントが絞られ，効率的な文書レビューと，仕様書の品質向上が期待される．

更新されたプログラム仕様書は，define.xml を作成するためのデータソースとなる．

4.4 プログラミングと品質保証

解析帳票は，アウトプットカタログと標準 DB を元に作成する解析帳票作成プログラムライブラリにあるものを用いて作成する．品質の保証されたプログラムを用いて解析帳票を作成するため，プログラム開発とその品質保証に係る時間を大幅に短縮できる．また，解析要件に応じて変更・追加した箇所を，特に注意すべき箇所として品質保証することで，効率的に品質保証できることが期待される。

5. まとめ

このように，標準を導入し，その恩恵を享受するには，まず導入する標準を元にプロセスを再構築し，標準を維持・管理していくための仕組みが必要だと考える．

当社では，4 章で紹介した統計解析業務プロセスの導入に着手したばかりだが，一部の工程は実装可能な段階まで来ているものもある。今後も，本業務プロセスの実装方法と成果，発生した問題点を共有していきたい．

参考文献

Michael J. Klepper and Barton Cobert (2011), Drug Safety Data : How to Analyze, Summarize, and Interpret to Determine Risk.

SAS を用いた EMICM アルゴリズムによる MST 推定の性能評価

中川 雄貴, 若林 将史, 浜田 知久馬
東京理科大学 工学研究科

Performance evaluation of the MST estimation method using the EMICM algorithm with SAS Software

Yuki Nakagawa, Masashi Wakabayashi, Chikuma Hamada
Graduate School of Engineering, Tokyo University of Science

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

SAS Global Forum 2010 において, 区間打切りデータにおける生存関数の推定をSAS で実行するプログラムが発表された[8]. 本発表ではその論文で用いられている EMICM アルゴリズムによる生存関数推定法の性能を検討する.

キーワード: EMICMアルゴリズム, MST, 区間打切り, 生存関数

生存時間解析

➢生存時間解析
試験終了時の患者の生存や追跡不能等の打切りを扱う

➢生存関数
ある時点tまでイベントが起きない確率 $S(t) = \Pr(T \geq t)$

➢本研究では生存関数に指数分布を仮定

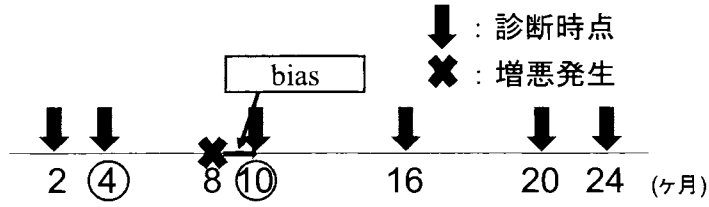
$h(t) = \lambda$
$S(t) = \exp(-\lambda t)$

区間打切りデータ

➢ある区間内でイベントが起きたことしかわからないデータ

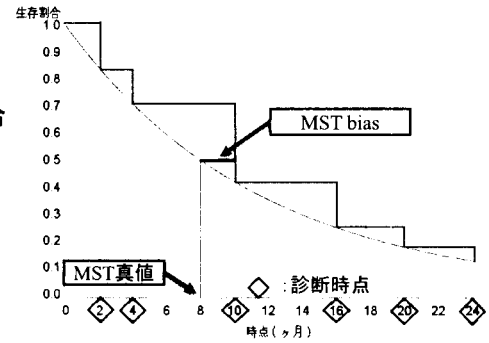
➢がんの臨床試験における区間打切りデータ
➢増悪は隣接した診断時点間で発生→正確な発生時点が確認できない
➢増悪確認時点を増悪発生時点として解析→生存時間の過大評価

➢区間打切りの例



MST : Median Survival Time

➢生存割合が 50% となる時点：がんの治療効果の指標
治療後の生存期間中央値 (Median Survival Time)
→増悪の 50% 時点



➢MST
指数分布を仮定した場合

$$S(t) = \exp(-\lambda t)$$
$$0.5 = \exp(-\lambda\ \mathrm{MST})$$
$$\mathrm{MST} = -\frac{\log(0.5)}{\lambda}$$

MSE : Mean Square Error

➢推定値のばらつきの指標
二乗損失の期待値：平均二乗誤差(Mean Square Error)

➢定義式

$$\begin{aligned}\mathrm{MSE} &= E\left[\{X-\mu\}^2\right] \\ &= E[\{(X-\theta)+(\theta-\mu)\}^2] \\ &= V[X]+(\theta-\mu)^2\end{aligned}$$

- X：推定値
- μ：真値
- θ：推定値の期待値$E[X]$

➢MSE を小さくする推定法→精度の高い推定法

カプラン・マイヤー法

- 生存関数の推定法
  生存時間分布に特定の分布の仮定はしない，打切り考慮

- カプラン・マイヤー推定量
  - t_i：イベント発生時点
  - n_i：リスク集合の大きさ
  - d_i：イベント発生数

$$\hat{S}(t) = \left(1 - \frac{d_1}{n_1}\right) \times \left(1 - \frac{d_2}{n_2}\right) \times \cdots \left(1 - \frac{d_i}{n_i}\right)$$
$$= \prod_{(t_i<t)} \left(1 - \frac{d_i}{n_i}\right)$$

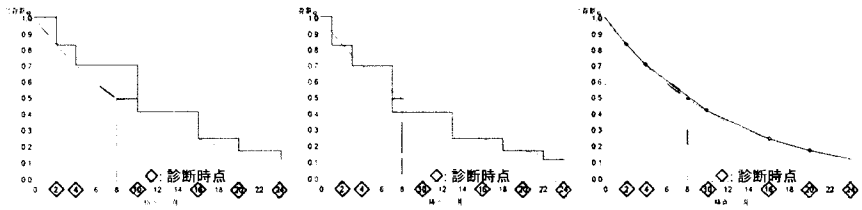
医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

区間打切りの場合の生存関数推定法

- カプラン・マイヤー法
  - 右側近似
    イベント確認時点をイベント発生時点として
    生存関数の推定を行う手法
  - 中点近似
    イベント確認時点と直前の診断時点の中点を
    イベント発生時点として生存関数の推定を行う手法

- EMICM(Expectation-Maximization Iterative Convex Minorant)
  アルゴリズム[4][5][6]
  生存時間分布に特定の分布を仮定することなく，
  区間打切りの影響を考慮して生存関数の推定を行う手法
  最尤推定で特定の基準を満たすまで推定値の更新を行う

生存関数推定法に基づくMST biasの例



右側近似　　中点近似　　EMICM

➢推定法が異なれば，MST biasの大きさも異なる

研究目的

➢生存関数推定法の性能をシミュレーションによって評価
  ➢複数のMSTの真値を設定し，複数の診断スケジュールにおいて総合的に性能がよい適切な生存関数推定法の検討
  ➢性能評価指標
    正確度：MST bias
    正確度＋精度：MSE

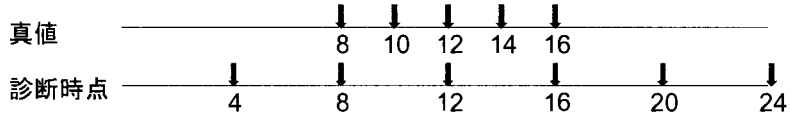
⇒臨床試験における適切なMSTの推定法の検討

シミュレーション条件

- 想定する条件
  - 想定する臨床試験：追跡期間 24 ヶ月の臨床試験
  - 生存時間分布：指数分布
  - 患者数：99 人
  - 診断回数：6 回
  - MST の真値：8, 10, 12, 14, 16 ヶ月
  - シミュレーション回数：10000 回

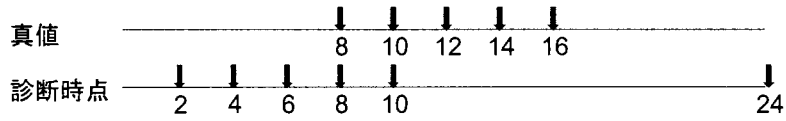
- 診断スケジュール
  - i (等間隔) ：4 , 8 , 12 , 16 , 20 , 24 ヶ月
  - ii (前半は密 後半は疎) ：2 , 4 , 6 , 8 , 10 , 24 ヶ月
  - iii (前半は疎 後半は密) ：2 , 16 , 18 , 20 , 22 , 24 ヶ月
  - iv (開始時と終了時に密) ：2 , 4 , 6 , 20 , 22 , 24 ヶ月

等間隔な診断スケジュール



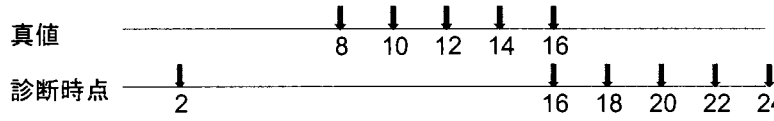
i	MST bias (月)			MSE (月^2)		
真値(月)	右側近似	中点近似	EMICM	右側近似	中点近似	EMICM
8	2.032	0.032	0.143	8.186	4.057	1.204
10	2.103	0.103	0.152	7.104	2.691	2.011
12	2.082	0.082	0.146	9.043	4.713	2.846
14	2.125	0.125	0.144	9.894	5.397	3.985
16	2.116	0.119	0.139	11.139	6.725	5.175

前半は密で後半は疎な診断スケジュール



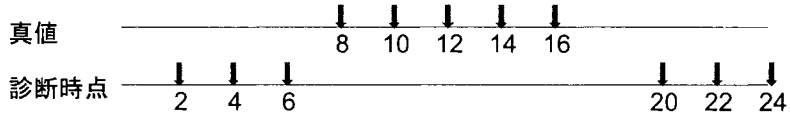
ii	MST bias (月)			MSE (月^2)		
真値(月)	右側近似	中点近似	EMICM	右側近似	中点近似	EMICM
8	1.696	0.377	0.084	15.552	5.499	1.385
10	6.944	2.907	0.278	99.504	25.872	2.805
12	10.436	4.103	0.661	128.492	23.248	4.273
14	9.799	2.886	0.937	98.800	9.240	5.090
16	7.981	1.002	1.024	63.963	1.172	5.482

前半は疎で後半は密な診断スケジュール



iii	MST bias (月)			MSE (月^2)		
真値(月)	右側近似	中点近似	EMICM	右側近似	中点近似	EMICM
8	8.000	1.000	2.082	64.000	1.000	4.967
10	6.000	−0.996	1.626	36.011	1.019	3.611
12	4.040	−2.850	1.151	16.418	9.337	2.837
14	2.424	−3.534	0.678	6.906	22.895	2.907
16	1.549	−2.412	0.351	5.817	27.448	4.157

開始時と終了時に密な診断スケジュール



iv	MST bias (月)			MSE (月^2)		
真値(月)	右側近似	中点近似	EMICM	右側近似	中点近似	EMICM
8	11.574	4.756	1.069	139.744	24.513	3.215
10	9.993	2.996	1.402	99.958	9.008	3.908
12	8.000	1.000	1.382	64.003	1.008	4.055
14	6.011	−0.956	1.198	36.170	1.275	4.058
16	4.124	−2.554	0.915	17.321	10.099	4.209

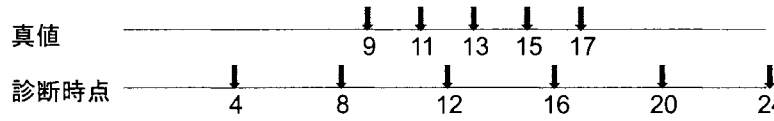
等間隔な診断スケジュールにおける中点近似の性能評価に対する再検討

- 等間隔な診断スケジュールにおいて中点近似が MST bias 最小
  - →真値の設定がたまたま中点近似の性能を上げた可能性
    以下の設定でシミュレーションを実行，性能の再検討

- 想定する条件
  - 想定する臨床試験：追跡期間24ヶ月の臨床試験
  - 患者数：99 人
  - MST の真値：9, 11, 13, 15, 17 ヶ月
  - シミュレーション回数：10000 回

- 診断スケジュール
  - i (等間隔)：4 , 8 , 12 , 16 , 20 , 24 ヶ月

等間隔な診断スケジュール



i	MST bias (月)			MSE (月^2)		
真値(月)	右側近似	中点近似	EMICM	右側近似	中点近似	EMICM
9	2.244	0.244	**0.153**	8.108	3.132	**1.619**
11	2.047	**0.047**	0.165	7.916	3.725	**2.360**
13	2.120	**0.120**	0.163	9.561	5.078	**3.348**
15	2.138	**0.138**	0.165	10.611	6.057	**4.498**
17	2.126	**0.136**	0.173	11.626	7.242	**5.737**

まとめ

- シミュレーションによって生存関数推定法の性能を評価
  - 診断時点が等間隔な場合は中点近似が MST bias 最小
    →EMICM アルゴリズムによる推定も大きな差はない
  - すべての診断スケジュールにおいて
    EMICM アルゴリズムによる推定は安定して性能がよい

⇒EMICM アルゴリズムによる推定を用いるのが好ましい

参考文献

[1] Andrzej P., Marek K.. Bioinformatics. Springer. 2007.
[2]Dempster A.P., Laird N.M., Rubin D.B.. Maximum likelihood from incomplete data via The EM algorithm. *Journalo f theRoyalS tatisticalS ociety*, Series B, 39, 1-38. 1977.
[3] Efron B.. The two sample problem with censored data. In Proc. *5th Berkeley Symp.* on Math. Statist. Prob. . Berkeley : University of California Press, 831-853. 1967.
[4] Groeneboom P., Wellner J.A.. Information bounds and nonparametric maximum likelihood estimation. *DMV Seminar*, Band 19, Birkhauser, New York. 1992
[5] Jianguo S.. The Statistical Analysis of Interval-censored Failure Time Data. Springer. 2006.

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

参考文献

[6]Turnbull B. W.. Nonparametric estimation of survivorship function with doubly censored data. *Journal of the American Statistical Association*, 69, 169-173. 1974.
[7] Wellner J. A., Zhan Y. A hybrid Algorithm for Computation of the Nonparametric Maximum Likelihood Estimator from Censored Data, *Journal of the American Statistical Association,* 92, 945-959. 1997.
[8] Ying S., Gordon J., Se H. K.. Analyzing Interval-Censored Survival Data with SAS Software. *SAS Global Forum 2010*, paper257. 2010.
[9]大橋靖雄, 浜田知久馬. 生存時間解析. 東京大学出版会. 2010

アジェンダ

1. データ公開・共有の背景
2. データ公開・共有の概要
3. SASの役割
4. CTDTソリューションの概要
5. Multi Sponsor Environment（MSE）
6. CTDTにおけるホットトピック

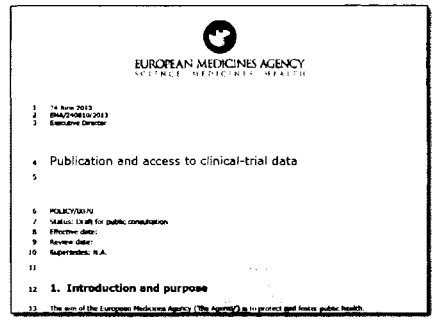
1. データ公開・共有の背景

データ公開・共有の背景　　背景

- 製薬企業が、NDA(新薬承認申請)を行う際に、
  データなどのエビデンスを、
  FDA、EMAやPMDAなどの
  規制当局に提出
- それらのデータは基本的には、
  提出する製薬企業、及び、
  受領する規制当局以外の
  第三者は入手不可
- 医薬品業界は、
  この従来より独占権を持って扱ってきた
  データの管理について、考え方を変えつつある状況

データ公開・共有の背景　EMAのドラフトガイドライン(1/2)

EMA(欧州医薬品庁)が、2014年6月24日にリリースした、データ公開・共有に関するドラフトポリシーにて患者レベルデータの公開などがうたわれており、医薬品業界関係者の関心が高まりました。



EUROPEAN MEDICINES AGENCY

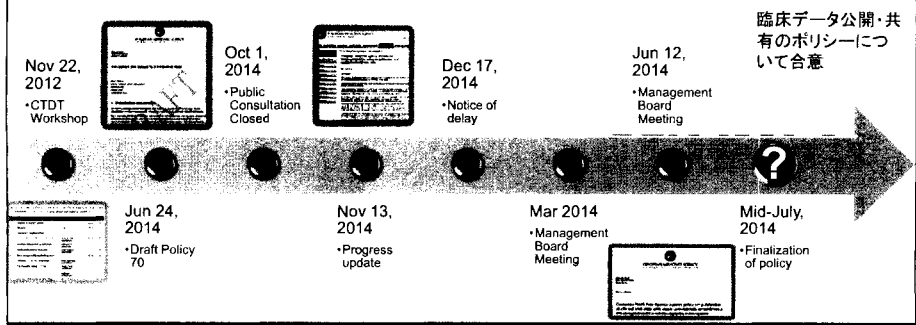
Publication and access to clinical-trial data

1. Introduction and purpose

Strictly Confidential URL: http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/06/WC500144730.pdf

データ公開・共有の背景　EMAのドラフトガイドライン(2/2)

多数の、意見の異なるパブリックオピニオンの受領などを通じて、EMAの意思決定は遅延してきましたが、データ公開・共有のポリシーに関して6月に方向性について合意し、最終化中の詳細手順を含め、7月中旬に完成する見込みです。

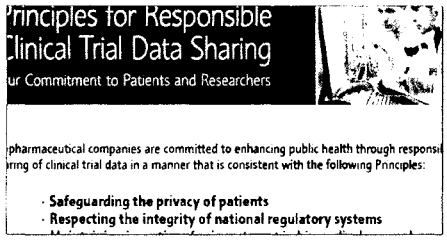


Strictly Confidential

データ公開・共有の背景　　efpiaとPhRMA共同のコミットメント

2013年7月、欧州製薬業団体連合会(efpia)と米国研究製薬工業協会(PhRMA)が、データ公開・共有に関して、共同でコミットメントを行ないました。

efpia　PhRMA



5つのコミットメント

1. 患者レベルデータの研究者との共有の推進
2. 臨床試験情報(最低でもCSRの概要)への市民のアクセスの推進
3. 臨床試験に参加した患者への結果の共有
4. 臨床試験情報を共有する手続きの保証
5. 臨床試験の結果の公開に関するコミットメントの再確認

2. データ公開・共有の概要

データ公開・共有の概要　Who, what, why and How?(1/2)

データ公開・共有とは何か？

- ライフサイエンス企業が、匿名化された患者レベルの臨床試験データへのアクセスを、正当な研究目的のために、提供すること

誰が提供するのか？

- ライフサイエンス企業
- （非営利団体を含む、その他の団体の可能性もあり）

誰がそれを使うのか？

- 業界や学会の研究者（スポンサーではない）
- （将来的には、医療の利用者、CRO等も可能性あり）

どのように使うのか？

- 独立した委員会（Independent Panel）が定義する、「正当な」研究目的
- 一般的な探索のためではなく、研究に関する提案書の提出が必須
- ライフサイエンス企業は、結果にアクセス

データ公開・共有の概要　Who, what, why and How?(2/2)

誰が関心を持っているのか？

- 現在、もしくは将来ソリューションを評価中の製薬企業
- 障害となる規制への対処に関心を持つ、企業
- 付加価値サービスを検討する第三者

どのような情報が含まれるのか？

- 患者レベルの臨床試験データ
- 関連文書

何故行なうのか？

- 結果の再現、検証を可能とすること
- 新たな知見を生むための、貴重なリソース
- 治験参加者への義務、彼らのデータの再利用
- 公衆衛生の改善、患者の安全性の改善、医薬品開発の促進
- 世間からの信頼の醸成
- 潜在的な規制に対する積極的な対応

3. SASの役割

SASの役割　　**取り組み経緯**

SASは、2013年前半に先行企業様のデータ公開、共有環境を構築した後、先行企業様のお力をお借りしながら、本テーマに関するミーティングなどをコーディネートして参りました。

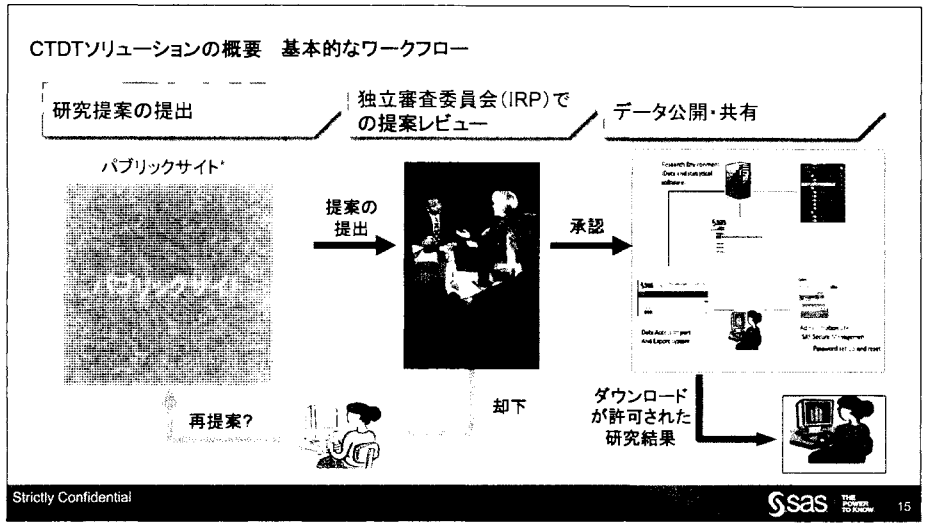
2013年前半	先行企業様の環境構築（臨床データレポジトリ、SDD*をベースに開発）
2013年10月17日	第1回CTDTフォーラム開催（SAS米国本社）
2013年12月	第1回CTDT日本セミナー開催（東京、大阪　二回開催）
2014年2月11日	第2回CTDTフォーラム開催（SAS米国本社）
2014年第一四半期	「SAS® Clinical Trial Data Transparency」を製品化
2014年4月29日	第3回CTDTフォーラム開催（SAS英国オフィス）
2014年第二四半期	Multi Sponsor Environment（MSE：後述）を提供開始
2014年8月下旬	第2回日本CTDTセミナー開催（予定）

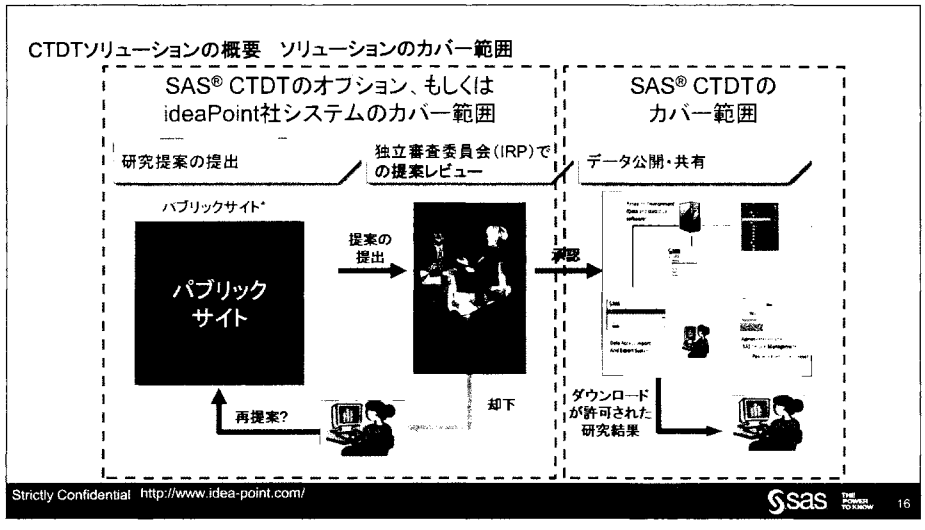
SASの役割　　ご参考資料

ミーティング等の様子は、以下のリンクからご参照頂くことが可能です。(一部、登録が必要)
先進企業の皆様が、CTDTの取り組みの意図やオペレーションについて共有されております。

- 第1回CTDTフォーラムのビデオ、マテリアル
  - http://www.sas.com/en_us/offers/13q4/Clinical-Trials-Data-Transparency-Forum-2286954/register.html
- 第2回CTDTフォーラムのビデオ、マテリアル
  - http://www.sas.com/en_us/offers/14q1/Clinical-Trials-Data-Transparency-Forum-Post-Event-2291234/register.html
- 第3回CTDTフォーラム
  - http://www.sas.com/apps/webnet/event_show_video.html?playlist=3449507691001&index=ev_ct2014 ※ビデオ
  - http://www.sas.com/offices/europe/uk/downloads/clinicaltrialdata-exec-summary.pdf ※イベントサマリー
- エグゼクティブに対するビデオインタビュー
  - http://www.sas.com/apps/webnet/event_show_video.html?playlist=3468745410001&index=ev_ctdtViews2014

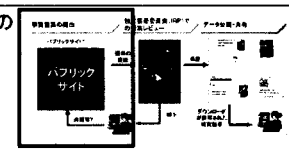
4. CTDTソリューションの概要


CTDTソリューションの概要　基本的なワークフロー
研究提案の提出
独立審査委員会(IRP)での提案レビュー
データ公開・共有
パブリックサイト*
提案の提出
承認
再提案?
却下
ダウンロードが許可された研究結果
Strictly Confidential
sas
15


CTDTソリューションの概要　ソリューションのカバー範囲
SAS® CTDTのオプション、もしくはideaPoint社システムのカバー範囲
SAS® CTDTのカバー範囲
研究提案の提出
独立審査委員会(IRP)での提案レビュー
データ公開・共有
パブリックサイト*
パブリックサイト
提案の提出
承認
再提案?
却下
ダウンロードが許可された研究結果
Strictly Confidential http://www.idea-point.com/
sas
16

CTDTソリューションの概要　基本的な要件

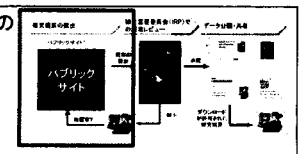
研究提案の提出



- パブリックサイトでの情報掲載
  - 公開可能な情報(プロトコル情報、患者レベルデータ等)の説明
  - 公開可能な試験のリスト
  - 研究提案プロセスの説明
  - データ利用同意書(Data Usage Agreement)の説明
  - 匿名化の方法の説明
  - 当該時点までの研究提案数、その承認数等の指標、等

CTDTソリューションの概要　基本的な要件

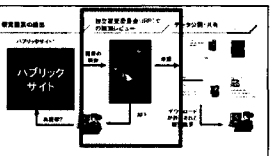
研究提案の提出



- 研究提案のハンドリングの為の機能
  - 研究提案の入力、受領
  - 独立審査委員会(IRP)への通知

CTDTソリューションの概要　基本的な要件

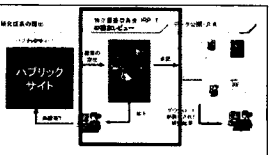
独立審査委員会(IRP)での提案レビュー



- 独立審査委員会(IRP)での提案レビュー
  - 新たにレビューが必要な研究提案の通知
  - 研究提案へのアクセス
  - スポンサーから追加情報が必要な場合等の、スポンサーへのアップデート
  - 結果のスポンサーへの通知

CTDTソリューションの概要　基本的な要件

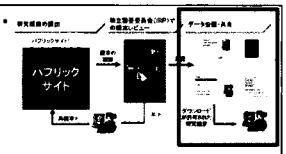
独立審査委員会(IRP)での提案レビュー



- 独立審査委員会(IRP)での提案レビュー
  - 【承認の場合】
    - 必要なデータのスポンサーへの通知
    - スポンサー:新規データの場合、匿名化
    - システム:リサーチャー向けの研究領域とアカウントの用意
    - システム:新規データの場合、データのロード
    - システム:研究者とデータの関連付け
    - 研究者に、承認された旨と初期ログイン情報の通知
  - 【却下の場合】
    - 研究者に、却下とその理由の通知

CTDTソリューションの概要　基本的な要件

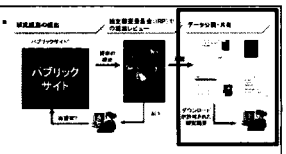
データ公開・共有



- データ公開・共有：情報
  - プロジェクト領域　※クラウド上の領域
    - 形式を問わず、ファイルを保管
      （sasプログラム、sasデータセット、pdfファイル、wordファイル・・・）
    - データや情報へのアクセスと閲覧
    - 新たな情報のインポート　※スポンサーが自動もしくは手動でコントロール可能
    - 情報のエクスポート　※スポンサーが自動もしくは手動でコントロール可能
  - 分析ツールへのアクセス
  - サードパーティツール（例：R）へのアクセス

CTDTソリューションの概要　基本的な要件

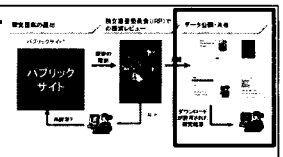
データ公開・共有



- データ公開・共有：分析
  - 【SAS】　※クラウド上の領域
    - SASプログラムの開発、実行環境（ログやリスト、アウトプットの閲覧含む）
    - 版管理
    - 他の研究者とのデータ共有
  - 【Non-SAS】　※クラウド上の領域
    - SAS環境（SAS-IML）を通じたRへのアクセス
    - Rへの直接アクセス

CTDTソリューションの概要　基本的な要件

データ公開・共有

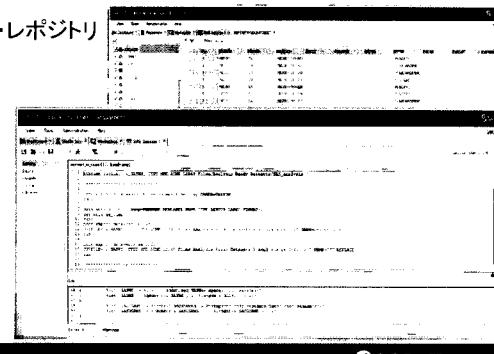


- サポートと管理
  - ホスティングとアドミニストレーション
    - トレーニング（研究者向けのトレーニングビデオを用意）
    - セルフヘルプ
    - バックアップ、システムのモニタリング
    - サービスレベルアグリーメント（SLA）
    - ソリューションや環境のアップデート
    - テクニカルサポート　※弊社米国本社にて集中対応
    - 分析サポート　※スポンサーが担当

CTDTソリューションの概要　セキュアな臨床分析プラットフォーム

CTDTは、CFR 21 Part 11準拠やグローバル開発の効率化目的で実績のある
臨床データレポジトリ（CDR）、SAS® Drug Developmentをベースに開発されております。

- パーミッション・ベースのファイル・レポジトリ
- ファイルの版管理
- 自動の監査証跡取得機能
- データ・ブラウザ
- SASプログラム開発、実行機能
  - プログラム開発環境の統合
    （従来のSASプログラマーが
    馴染みやすい環境）
  - 実行環境
- 結果のトレース機能（manifest）
- 必要に応じて、外部の分析環境
  との統合



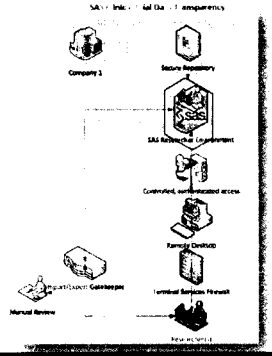
5. Multi Sponsor Environment(MSE)

Strictly Confidential

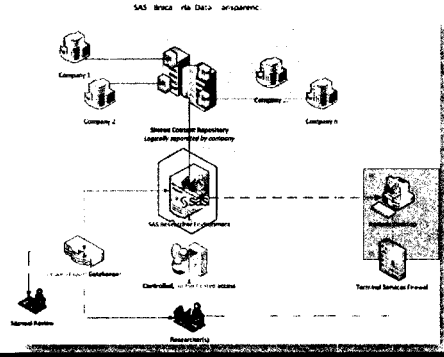
sas

Multi Sponsor Environment(MSE)　アプローチの比較

Single Instance(従来のアプローチ)



Multi-Sponsor Environment



Strictly Confidential

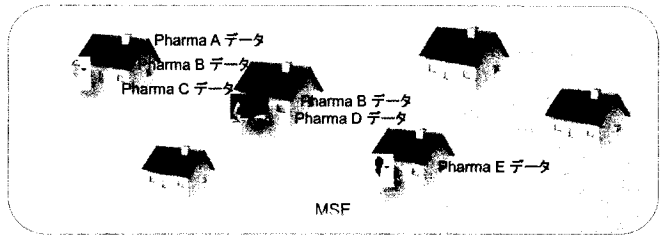
sas THE POWER TO KNOW 26

Multi Sponsor Environment(MSE)

MSEのイメージ

各研究者チームは、それぞれの研究プロジェクトの為に、一軒の「家」を使用

- いくつのスポンサーのデータにアクセスするかは研究者の自由(但し、承認されれば)
- 研究者チームが複数のプロジェクトを承認された場合は、複数の「家」を使用



Multi Sponsor Environment(MSE)

MSEのメリットとデメリット

Pros

- 研究者は、一つのプロジェクトにおいて、複数のスポンサーのデータにアクセスすることも可能
- スポンサーはホスティング費用を分担することが可能
- 臨床データの公開、共有について、業界のデファクト・スタンダードを確立可能
- 環境構築(go-live)までの期間短縮が可能(数か月から数週間に)

Cons

- スポンサー毎に、設定やブランディングできる範囲の制約
- 物理的には、情報を他のスポンサーのデータと一緒に保管
- 他のスポンサーと同一のソフト、ツール、プロセスを適用することが必要
- 他のスポンサーとの法的な合意が必要


sas

生存時間解析におけるノンパラメトリック検定の多重比較に関する研究

島村 文也 ・ 浜田 知久馬 ・ 佐野 雅隆
東京理科大学大学院 工学研究科 経営工学専攻

背景・目的

➢ 生存時間解析における多群比較[3]

- 医薬研究では，イベント発生までの時間（生存時間）を評価
  - –ノンパラメトリック検定を用いて，各群の生存関数を比較

➢ ノンパラメトリック検定の多群比較における問題点

- 検定の多重性[1]
  - –検定を複数回行うと，α エラーが増大する
  - –α エラーを有意水準以下に抑えるために，厳しめに検定を行う多重比較法が複数存在
    - 生存関数の多群比較における多重比較法の性能評価は十分でない
- 生存時間に対する順位付けの方法の決定
  - –**Joint-ranking 法**
    - すべての群の生存時間に対して順位付けを行う
  - –**Separate-ranking 法**
    - 比較する 2 つの群の生存時間に対してのみ順位付けを行う

e.g.,
DOSE 10 と DOSE 30 の各群の生存関数を比較

DOSE	生存時間（週）				
10	27	31	32	33	34
30	21	23	28	29	30
90	20	22	24	25	28

➢DOSE 90 の群の生存時間を考慮

Joint-ranking 法

DOSE	順位				
10	8	12	13	14	15
30	2	4	7	10	11
90	1	3	5	8	9

Separate-ranking 法

DOSE	順位				
10	4	7	8	9	10
30	1	2	3	5	6
90					

➢生存時間解析において，より性能の高い多群比較を行うための適切な方法を検討
- Joint-ranking 法 と Separate-ranking 法 の性能評価
- α エラーを有意水準付近以下に抑えつつ，検出力の高い多重比較法の検討

方法

➢ 本研究で検討する多重比較法[1]

- Bonferroni 法

$$\alpha' = \frac{\alpha}{h}$$

- Sidak 法

$$\alpha' = 1 - (1-\alpha)^{\frac{1}{h}}$$

μ : 母数
α : 設定する有意水準
h : 試験全体での仮説数
α' : 各比較に用いる有意水準
r : 群の数
i, j : 群

- Holm 法
  - –p 値を小さい順に並べてから設定した有意水準 α を調整

$$p_1 \le p_2 \le \cdots \le p_h \quad \rightarrow \quad \alpha'_1 = \frac{\alpha}{h},\ \alpha'_2 = \frac{\alpha}{h-1},\ \cdots,\ \alpha'_h = \frac{\alpha}{1}$$

- Scheffe 法
  - –自由度を $r-1$ とした χ^2 分布を仮定することで調整

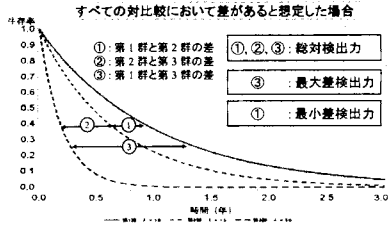
$$\text{non-adjusted} : p = \Pr(\chi_i^2 > \chi_{ij}^2)$$
$$\text{Scheffe} \quad : p = \Pr(\chi_{r-1}^2 > \chi_{ij}^2)$$

- Tukey 法
  - –χ^2 分布の代わりに，「student 化された範囲の q 分布」を用いて調整

➢ 多重比較法における検出力[2]

- 多重比較では複数個の対立仮説があるため，検出力の定義も様々
  - –それぞれの定義に応じて各手法の優劣の程度も変化

すべての対立比較において差があると想定した場合



シミュレーション条件

- 生存時間分布 : 指数分布
- 各群の症例数 : 50 例 (3 群比較)
- 検定方法 : ログランク検定
- 検討した方法 (計 10 通り)

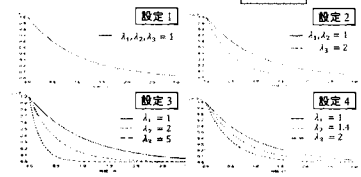
{ Joint-ranking 法 / Separate-ranking 法 } × { Bonferroni 法, Sidak 法, Holm 法, Scheffe 法, Tukey 法 }

- シミュレーション回数 : 10,000 回
- 有意水準 : 両側 5 %
- 評価指標 : α エラー，総対検出力，最大差検出力，最小差検出力

➢生存関数におけるハザードの設定[2]

i : 群
λ : ハザード

- $S_i(t) = \exp(-\lambda_i t)$ $[i = 1, 2, 3]$



結果・考察

α エラー（両側 5 %）

多重比較法	設定 1		設定 2	
	Joint	Separate	Joint	Separate
Bonferroni	0.0465	0.0470	0.0179	0.0171
Sidak	0.0471	0.0479	0.0182	0.0178
Holm	0.0465	0.0470	0.0356	0.0388
Scheffe				
Tukey	0.0529	0.0537	0.0214	0.0200
調整なし	0.1239	0.1267	0.0436	0.0504

設定 1 に関して，Tukey 法が多少有意水準を上回る
- –α エラーを有意水準以下に抑えられないログランク検定の性質が原因[4]

総対検出力

多重比較法	設定 2		設定 3		設定 4	
	Joint	Separate	Joint	Separate	Joint	Separate
Bonferroni	0.8521	0.7351	0.8585	0.8172	0.0040	0.0100
Sidak	0.8546	0.7402	0.8614	0.8184	0.0040	0.0100
Holm						
Scheffe	0.6755	0.7108	0.6258	0.7993	0.0028	0.0182
Tukey	0.8273	0.7503	0.8906	0.8325	0.0048	0.0225
調整なし	0.8247	0.8002	0.8625	0.9100	0.0437	0.0911

Separate-ranking 法と Holm 法を用いた場合が最も高い

最大差検出力

多重比較法	設定 2		設定 3		設定 4	
	Joint	Separate	Joint	Separate	Joint	Separate
Bonferroni	0.7953		1.0000	1.0000	0.8437	0.8398
Sidak	0.7977		1.0000	1.0000	0.8452	0.8417
Holm			1.0000	1.0000	0.8498	0.8466
Scheffe	0.7773		1.0000	1.0000	0.8310	0.8247
Tukey	0.8123		1.0000	1.0000		
調整なし	0.9205		1.0000	1.0000	0.9292	0.9270

設定 4 に関しては，Tukey 法が最も検出力が良いが他の手法と大きな差はない

最小差検出力

多重比較法	設定 2		設定 3		設定 4	
	Joint	Separate	Joint	Separate	Joint	Separate
Bonferroni	0.7953	0.8398	0.8068	0.8382	0.2452	0.2305
Sidak	0.7977	0.8417	0.8078	0.8393	0.2471	0.2325
Holm						
Scheffe	0.7773	0.8247	0.7923	0.8232	0.2299	0.2147
Tukey	0.8123	0.8466	0.8210	0.8505	0.2607	0.2467
調整なし	0.9205	0.9270	0.9101	0.9256	0.3965	0.3844

設定 4 に関しては，Joint-ranking 法が全体的に 1~2 % 程上回ったが，Holm 法においては差はない

まとめ

➢生存時間に対する順位付けの方法に関して
- α エラーは各手法で名義水準以下に保たれている
- 全体的な検出力は Separate-ranking 法の方が高い
  - –設定 4 の最大差・最小差検出力に関してのみ joint-ranking法が高い

➢多重比較法に関して
- Scheffe 法は保守的（他の手法も問題なし）
- 検出力

Holm ＞ Tukey ＞ Bonferroni ≅ Sidak ＞ Scheffe

推奨する手法
Separate-ranking 法
Holm 法

参考文献

[1] 永田靖，吉田道弘．統計的多重比較法の基礎．サイエンティスト社．2001.

[2] 永田靖，松田直一．多重比較における新たな検出力の提案と各手法の特徴比較．応用統計学．19, 93-113. 1991.

[3] 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.

[4] A. M. Kellerer, D. Chmelevsky (1983). Small-Sample Properties of Censored-Data Rank Tests. *Biometrics*. 39. 675-682.

抗がん剤の第2相試験における被験者数変動を考慮した最適デザイン

豊泉 滋之 / ファイザー株式会社/臨床統計部

浜田 知久馬 / 東京理科大学

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

抗がん剤のスクリーニングを目的とした臨床試験は，効果を有するか否かを見極めることが最大の目的であり，効果が見込まれない薬剤候補を早期に判断できるよう，多段階（多くは2段階）に分けて効果を判定する試験デザインが用いられる．古典的な2段階デザインとして，平均症例数を最小化または最大症例数を最小化する試験デザイン（Simon法）が多用されている．Simon法では計画症例数と実際に解析対象となる症例数（実症例数）が一致する場合にその最適性が保障されるが，被験者の途中脱落や計画よりも多くの症例を登録することにより，計画症例数と実症例数が異なる場合に最適性が失われる問題が存在する．また，計画症例数と実症例数が異なる場合に，効果の有無の判定基準（境界例数）を実症例数に基づき再設定する手法が提案されているが，計画症例数からの変動幅に制限がなく，恣意性が入り易いといった問題がある．本発表では実症例数の変動を考慮し，計画症例数と実症例数が異なる場合においても最適性が保障される試験デザインを提案し，その試験デザインを求めるSASマクロについて発表する．

その他関連分野

SAS/IMLによる医療経済評価（モデル分析）

奥山ことば
MSD株式会社
グローバル研究開発本部 臨床研究統計部

Health Technology Assessment (Model Analysis) by Using SAS/IML

Kotoba Okuyama
Biostatistics & Research Decision Sciences (BARDS)
Japan Development, MSD K.K.

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨：

医療費の高騰から、医療経済学評価を用い、限られた財源を医療資源へ最適配分することも注目されている。そのモデル分析（決定樹、マルコフモデル等）の基礎的概念と、計算過程でIMLを利用する方法を紹介する。

キーワード：医療経済学、HTA、決定樹、マルコフモデル、費用対効果、ICER

発表内容

1. はじめに
2. 判断分析モデル
   決定樹、QALY、費用
3. マルコフモデル
   状態推移図、割引、結果解釈(ICER)、
   不確実性の考慮
4. マルコフモデル(シミュレーション)
5. 海外での医療経済評価
6. おわりに

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

はじめに

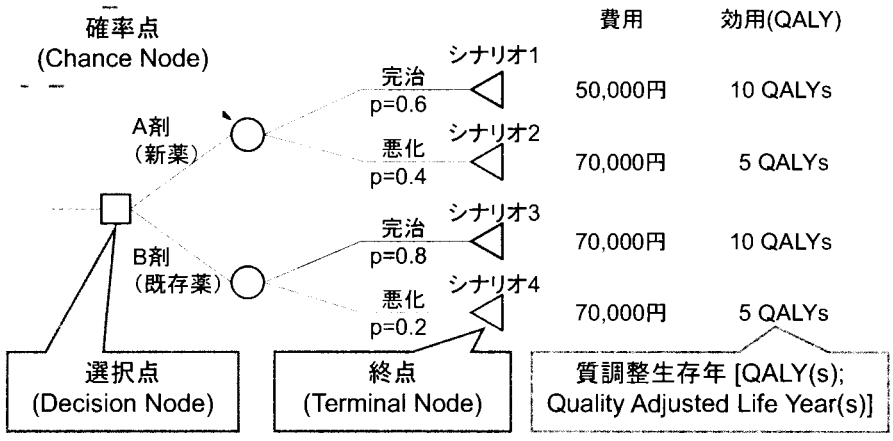
- 我が国の医療費は増加の一途
  - 平成24年度の国民医療費は38.4兆円(毎年、約1兆円増のペース)
  - 限られた財源の医療資源への最適配分又は効率的配分の重要性
- 業界の動き
  - 中医協の専門部会では、費用対効果評価を薬価や診療報酬に反映することについて検討を重ねている。
  - 昨年、厚生労働省研究班(福田班)から「医療経済評価研究における分析手法に関するガイドライン」が明らかにされた。

  →遠くない将来、既収載の薬価の見直し等を皮切りに、費用対効果評価が薬価収載時に求められる可能性がある。

- 医療経済学評価でのモデル分析手法:
  決定樹、マルコフモデル、シミュレーション(マルコフモデル)等

- 本発表では、手法の基礎的概念と、計算にSAS/IMLを利用する方法を紹介する。

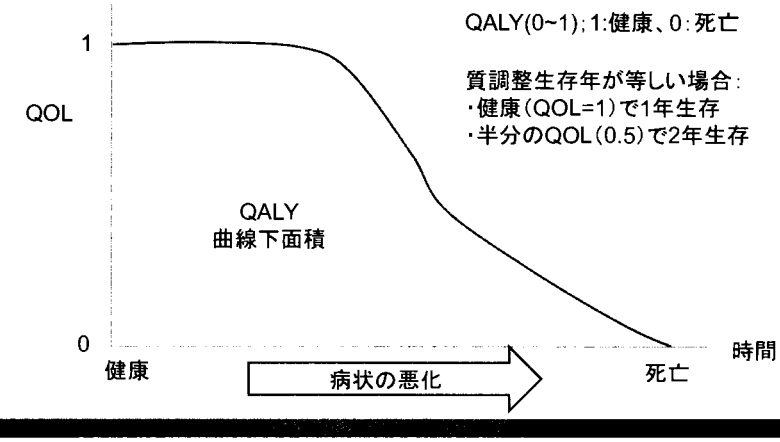
判断分析モデル(Decision Analysis Model)・判断樹/決断樹/決定樹(Decision Tree)



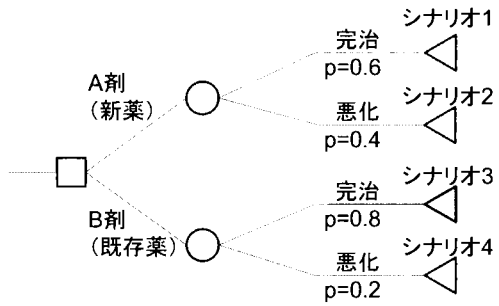
医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

質調整生存年;QALY(s)[Quality Adjusted Life Year(s)]

- 生存期間にQOLを乗じた指標



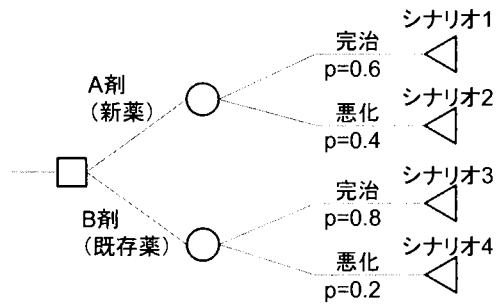
各薬剤の平均費用



	費用		期待費用
シナリオ1	50,000円	x0.6 =	30,000
シナリオ2	70,000円	x0.4 =	28,000
シナリオ3	70,000円	x0.8 =	56,000
シナリオ4	70,000円	x0.2 =	14,000

A剤の平均費用 = 30,000 + 28,000 = 58,000円
B剤の平均費用 = 56,000 + 14,000 = 70,000円

各薬剤の平均QALY



	QALY		期待QALY
シナリオ1	10 QALYs	x0.6 =	6 QALYs
シナリオ2	5 QALYs	x0.4 =	2 QALYs
シナリオ3	10 QALYs	x0.8 =	8 QALYs
シナリオ4	5 QALYs	x0.2 =	1 QALYs

A剤の平均QALY = 6 + 2 = 8 QALYs
B剤の平均QALY = 8 + 1 = 9 QALYs

各薬剤の費用対効果 及び増分費用対効果比(ICER; Incremental Cost-Effectiveness Ratio)

	費用 (円)	QALY	効果1単位 あたりの費用 (円/QALY)	費用差 (B-A)	QALY差 (B-A)	ICER 費用差/QALY差 (円/QALY)
A剤 (新薬)	58,000	8	7,250			
B剤 (既存薬)	70,000	9	7,778	12,000	1	12,000

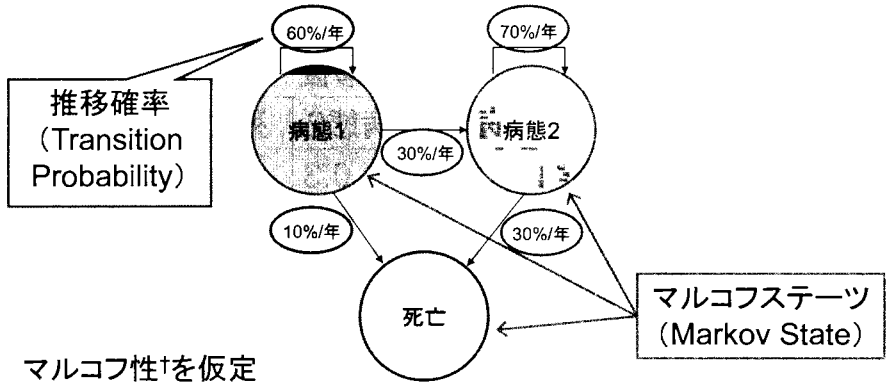
QALYあたりの費用

B剤を投与し、1QALY 増やすのに必要な費用

プログラム(decision_tree.sas)で出力(テキスト出力)する内容。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

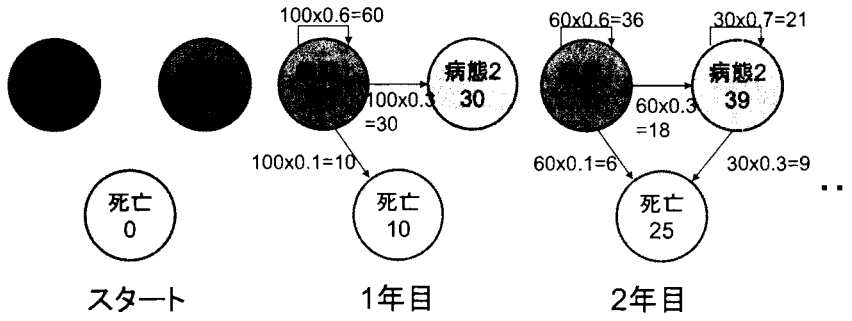
マルコフモデル 状態推移図(State Transition Diagram) 要因相関図(influence diagram)



マルコフ性†を仮定

† 現在の状態のみに依存して、将来の状態が確率的に決定

状態推移図のフロー



推移確率行列

- 推移確率を行列形式にまとめたもの。
- 群の数だけ推移確率行列も存在する。

		推移後		
		病態1	病態2	死亡
推移前	病態1	0.6	0.3	0.1
	病態2	0	0.7	0.3
	死亡	0	0	1

コホート分布の計算

1年目：
$$(100 \quad 0 \quad 0)\begin{pmatrix}0.6 & 0.3 & 0.1\\ 0 & 0.7 & 0.3\\ 0 & 0 & 1\end{pmatrix} = (60 \quad 30 \quad 10)$$

（病態1, 病態2, 死亡）

初期コホート分布

2年目：
$$(60 \quad 30 \quad 10)\begin{pmatrix}0.6 & 0.3 & 0.1\\ 0 & 0.7 & 0.3\\ 0 & 0 & 1\end{pmatrix} = (36 \quad 39 \quad 25)$$

⋮ ⋮

n年目：
$$(100 \quad 0 \quad 0)\begin{pmatrix}0.6 & 0.3 & 0.1\\ 0 & 0.7 & 0.3\\ 0 & 0 & 1\end{pmatrix}^{n}$$

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

費用・効果/効用ベクトル

	費用（万円/年）	効果/効用（QALYs）
病態1	5	0.9
病態2	10	0.5
死亡	0	0

新たな死亡のみ費用が発生する場合がある。

費用ベクトル

効果/効用ベクトル

費用の計算

開始年：
$$(100 \quad 0 \quad 0)\begin{pmatrix}5\\10\\0\end{pmatrix}=500$$

1年目：
$$(60 \quad 30 \quad 10)\begin{pmatrix}5\\10\\0\end{pmatrix}=600$$

2年目：
$$(36 \quad 39 \quad 25)\begin{pmatrix}5\\10\\0\end{pmatrix}=570$$

⋮ ⋮

n年目：
$$(100 \quad 0 \quad 0)\begin{pmatrix}0.6 & 0.3 & 0.1\\0 & 0.7 & 0.3\\0 & 0 & 1\end{pmatrix}^{n}\begin{pmatrix}5\\10\\0\end{pmatrix}$$

} 合計費用

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

割引計算

- インフレの考慮、時間選好等の考え方
- 将来の費用及び効用の両方を、現在の価値に換算する。
  t 年目の費用及び効用を $(1+\text{割引率})^t$ で割る。
- ガイダンスでは日本の経済実情を反映(2%)。
- 割引率を変化させた感度分析が必要(0～4%)。

	費用	現在の価値 (割引率2%)
開始年	100万円	100万円
1年目	100万円	100/1.02=98万円
2年目	100万円	$100/1.02^2$=96万円
合計	300万円	294万円

数式表現

- x_i: i年目のコホート分布のベクトル
- T: 推移確率行列
- c: 費用/効果ベクトル
- c_i: i年目の費用
- d: 割引率
- 各年の費用＝各年のコホート分布ベクトル×費用ベクトル

$$c_0 = x_0^T c$$
$$c_1 = x_1^T c = x_0^T Tc/(1+d)$$
$$c_2 = x_2^T c = x_0^T T^2 c/(1+d)^2$$
$$\vdots$$
$$c_n = x_n^T c = x_0^T T^n c/(1+d)^n$$

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

数式表現（つづき）

- 費用の合計

$$\sum_{i=0}^{n} c_i = \sum_{i=0}^{n} x_0^T T^i c/(1+d)^i$$
$$= x_0^T \left\{ I + \frac{1}{(1+d)} T + \frac{1}{(1+d)^2} T^2 + \cdots + \frac{1}{(1+d)^n} T^n \right\} c$$

QALYの計算も全く同様→IMLで実行可能

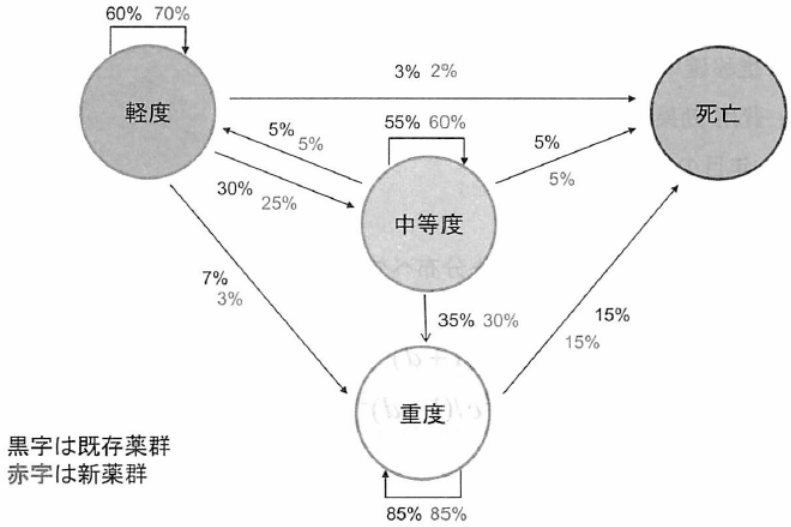
（注）$S=T/(1+d)$ とし、以下のようにも変形できるが、逆行列が一意に求まらないことが生じうるため、上記の計算式を使用する（そうすることで、経時的変化がトレース可能となる）。

$$\sum_{i=0}^{n} c_i = x_0^T \left\{ I + S + S^2 + \cdots + S^n \right\} c$$
$$= x_0^T (I - S^{n+1})(I - S)^{-1} c$$

$\because (I-A)^{-1} = I + A + A^2 + \cdots$ より

$$I + A + A^2 + \cdots + A^n$$
$$= (I-A)^{-1} - (A^{n+1} + A^{n+2} + \cdots)$$
$$= (I-A)^{-1} - A^{n+1}(I + A + A^2 + \cdots)$$
$$= (I-A)^{-1} - A^{n+1}(I-A)^{-1}$$
$$= (I - A^{n+1})(I-A)^{-1}$$

例：状態推移図



例：推移確率行列

		推移後 上段：A剤（新薬） 下段：B剤（既存薬）			
		軽度	中等度	重度	死亡
推移前	軽度	0.70 0.60	0.25 0.30	0.03 0.07	0.02 0.03
	中等度	0.05 0.05	0.60 0.55	0.30 0.35	0.05 0.05
	重度	0.0 0.0	0.0 0.0	0.85 0.85	0.15 0.15
	死亡	0.0 0.0	0.0 0.0	0.0 0.0	1.0 1.0

費用・効果/効用ベクトル

分析の期間(時間地平):20年
割引率:2%/年
開始時のコホート分布[100,0,0,0]

	費用(万円/年) 上段:A剤(新薬) 下段:B剤(既存薬)	QALYs
軽度	100 70	0.8
中等度	130 100	0.5
重度	180 150	0.1
死亡	0※ 0※	0.0
※新たな死に対する費用も考慮する場合がある。		

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

計算結果(Markov.sas)

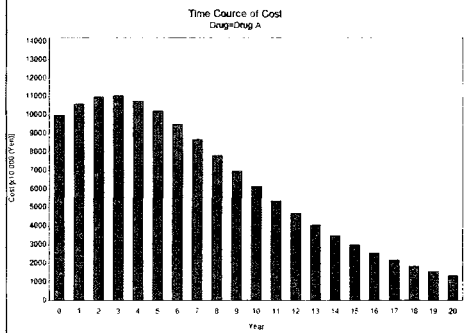
Cost-Effectiveness Assessment
Markov Model

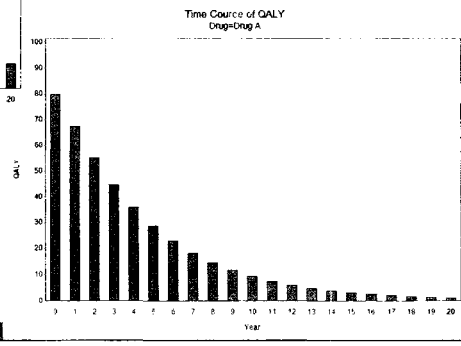
Total Cost for Drug A	Total QALY for Drug A	Total Cost for Drug B	Total QALY for Drug B	Cost-Benefit for Drug A	Cost-Benefit for Drug B
132871.57	423.773	99144.11	336.677	313.544	294.478

Difference in Cost (Drug A - B)	Difference in QALY (Drug A - B)	ICER
33727.46	87.0957	387.246

1QALYあたり388万円の費用増加だが、B剤(既存薬)からA剤(新薬)に変更することは、基準となる閾値の500万円/QALY※より小さいので、費用対効果が良いことになる。
※日本では500～600万円/QALYが目安とされている。

計算結果（Markov.sas）－つづき





医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

結果の解釈

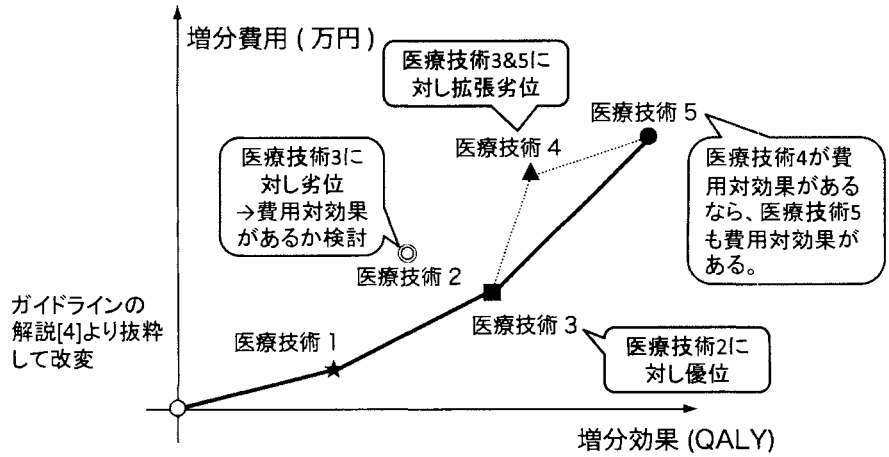
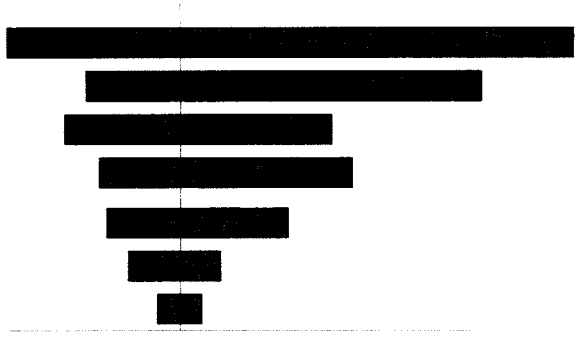


図4.1　劣位（医療技術２）と拡張劣位（医療技術４）

不確実性の考慮(一元感度分析)
トルネードチャート(竜巻図)

影響力の大きいパラメータから順番にプロット



薬剤の中止10%～50%

状態1から状態2への推移確率

状態3から状態4への推移確率

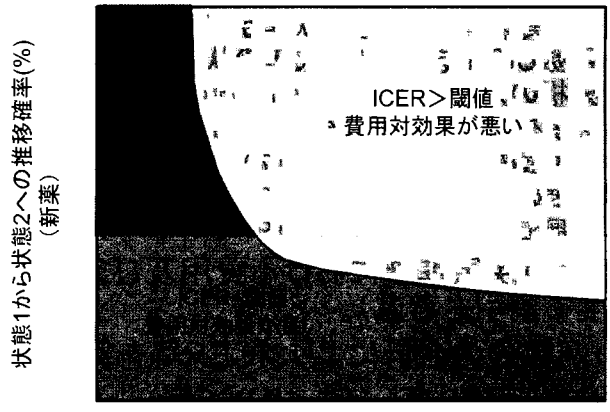
…

割引率 etc.

費用の増分／QALYの増分(実薬群－プラセボ群)

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

不確実性の考慮(二元感度分析)



コホート分布 vs モンテカルロシミュレーション

コホート分布

- 実行が可能で、パラメータ設定が明快。
- マルコフ性(メモリーレス)により、各患者が直前にどのステーツにいたのかが判別不能。
  - 複数のステーツが並行する場合(ステーツの履歴が重要)
  - ステーツ滞在時間より、推移確率が変わる場合
- 感度分析はパラメータの低次元(1~2次元)のみの評価で、ICER値は範囲で示され、尤度のような定量的に評価した値は算出できない。

モンテカルロシミュレーション(コホートではなく、患者単位でシミュレート)

- マルコフ性を超えた柔軟性を実現できる。
- 増分費用、増分効果、ICERを1つの値ではなく、分布として得るため、閾値を超える確率などを定量的に評価可能(=繰り返しが必要)。
- 確率的感度分析(PSA;Probabilistic sensitivity analysis)※に対応可能

  ※全パラメータに不確実性を導入する方法
- NICEがPSAを推奨したこともあり、シミュレーションが増加している。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

モンテカルロシミュレーション

- 疾患の進展を1例ずつ確率的にシミュレート
  - 前例で、実薬群で、現在の状態が「軽度」の場合
    一様乱数U(0,1)を発生し、次年の状態を決定し、死亡または分析の期間に達するまで繰り返す。

    軽度に留まる確率: 70%/年 (0≤U(0,1)<0.70)
    中等度に進展確率: 25%/年 (0.70≤U(0,1)<0.95)
    重度に進展確率: 3%/年 (0.95≤U(0,1)<0.98)
    死亡に進展確率: 2%/年 (0.98≤U(0,1) ≤1)
- 当該患者の疾患進展における費用及び効用を算出(各年度で割引も考慮)。
- 所与の例数分の合計により、ICER等を算出する。
- これを多数回繰り返し(例:10,000回)、平均値を推定値として用いる(10,000個のICERの平均値)

計算結果(Markov_patient_level.sas)

Cost-Effectiveness Assessment
Markov Model (Patient Level)

Total Cost for Drug A	Total QALY for Drug A	Total Cost for Drug B	Total QALY for Drug B	Cost-Benefit for Drug A	Cost-Benefit for Drug B
132841592.24	424034.70	99071545.78	336791.04	313.280	294.163

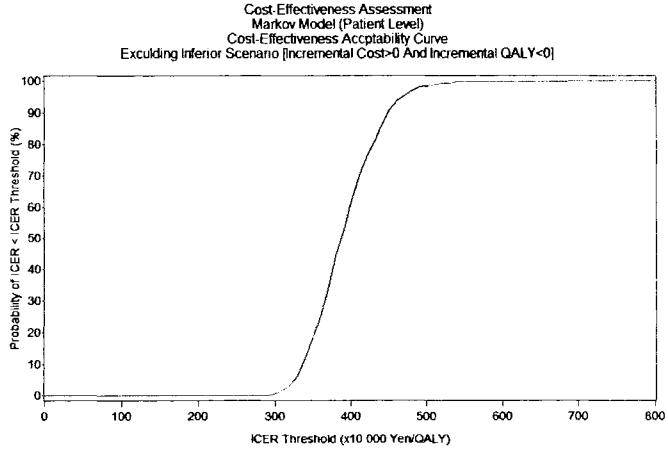
Difference in Cost (Drug A - B)	Difference in QALY (Drug A - B)	ICER
33770046.45	87243.67	387.077

上記は100,000例を1回だけシミュレートした結果で、コホート分布を利用したICER(387.246)とほぼ同じ。
100例を10,000回シミュレーション(平均ICER=417.830):バラツキ大
1000例を1000回シミュレーション(平均ICER=390.960)

より複雑なモデルは、繰り返し回数の増加が必要になる(上記の単純な例でも、1000例を10,000回シミュレーションは実行に90分弱)。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

計算結果(Markov_patient_level.sas)つづき
許容可能性曲線(acceptability curve)



QALYが減少し、且つ費用も増大する(劣位)のシナリオは除いて作成する。

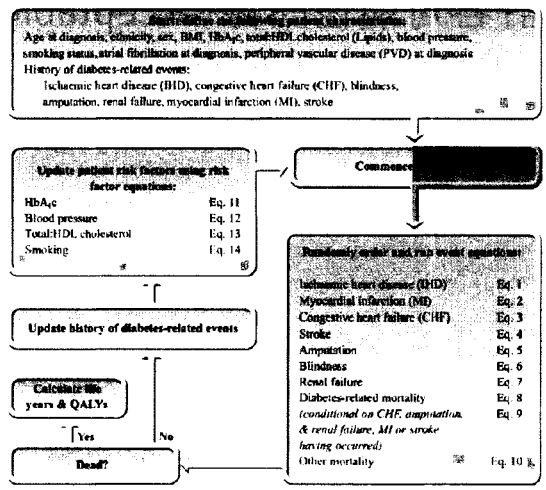
モンテカルロシミュレーション－つづき

- 症例の背景因子の違い、現在の状態、時間依存的に推移確率を変更することも対応可能
  - 背景の分布、時間依存的な推移確率等の追加情報が必要となるが、十分なエビデンスが存在しないかも。
  - 複雑なモデルでは計算時間が増大し、十分な感度分析の弊害となる可能性がある。
- コホート分布用いたモデルとの比較
  - 結果が同様であれば、コホート分布で十分かもしれない。
  - 比較した研究が少ないため、どちらが優れているか結論にまでは至っていない[11]

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

海外の医療経済学評価

Fig. 1. Algorithm for model simulation



UKPDS Outcomes Model [12]
(Clarke et al. 2004)より抜粋

おわりに

- 医療経済評価を通じて、限られた医療財源の効率的配分に寄与することも期待され、今後、医療経済評価が求められる可能性がある。
- 本発表では分析手法の基礎的概念と、SAS/IML等を利用して計算する方法を紹介した。
  - コホート分布かシミュレーションかは、取り組む問題の性質や複雑性などに依存する。
  - モンテカルロシミュレーションは柔軟性はあるが、実行時間が長い、追加情報の必要性の問題などがある。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

参考文献

[1]平成24年厚生労働科学研究費補助金 福田班, 医療経済評価研究における分析手法に関するガイドライン, Ver.1.0, 2013.

[2]白岩健, 福田敬, 五十嵐中, 池田俊也, CHEERS声明－医療経済評価における報告様式のガイダンス－, 保健医療科学, Vol.62, No.6, 641-666, 2013.

[3]白岩健, <総説>「医療経済評価研究における分析手法に関するガイドライン」の解説, 保健医療科学, Vol.62, No.6, 590-98, 2013.

[4]福田敬, 白岩健, 池田俊也 他, <解説>医療経済評価研究における分析手法に関するガイドライン, 保健医療科学, Vol.62, No.6, 625-40, 2013.

[5]白岩健, 福田敬, 五十嵐中, 池田俊也 訳, <解説>CHEERS声明--医療経済評価における報告様式のガイダンス－, 保健医療科学, Vol.62, No.6, 641-66, 2013.

[6]坂巻弘之, やさしく学ぶ 薬剤経済学, じほう, 2003.

[7]池上直己, 西村周三編著, 医療技術・医薬品, 勁草書房, 2005.

[8]鎌江伊三夫, 医薬経済学的手法による医療技術評価を考える<6>－データの不確実性をどう取り扱うか－, 医薬品医療機器レギュラトリーサイエンス, 44(1), 47-53, 2013.

[9]飯野四郎, 安田清美, 小林慎, 藤野志朗, 小川京子, C型慢性肝炎に対するIFN療法の費用効用分析－活動性投与と非活動性投与の比較, 日本医事新報, 3870, 10-15, 1998.

[10]池田俊也, 山田ゆかり, 池上直己, 抗痴呆薬ドネペジルの経済評価, 医療と社会, Vol.10, No.3, 2000.

[11]Briggs A, Claxton K, Sculpher M, Decision modeling for health economic evaluation, Oxford: Oxford University Press, 2006.

[12]Clarke P M, Gray A M, Briggs A, Farmer A J, Fenn P, Stevens R J, et al., A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model', Diabetologia, 47, 1747-59, 2004.

[13]小久保欣哉, 山田謙次, 医療技術評価(HTA:Health Technology Assessment)の政策動向と製薬企業における薬剤経済学の活用, 知的資産創造, 2013.

減らせ突然死
院外心肺停止のビッグデータから見えてくるもの

田久 浩志、田中 秀治
国士舘大学体育学部スポーツ医科学科

To Improve Resuscitation After Out-of-hospital Sudden Cardiac Arrest

TAKYU Hiroshi, TANAKA Hideharu

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

JMPを用いて92万件の病院前の心肺停止のデータを解析し、心臓突然死の基礎的背景を検討する。

キーワード:心肺停止、病院前、JMP 11.0

【はじめに】

- 我が国では2005-2012年にかけての92万件に及ぶ、院外心肺停止のデータベース（ウツタイン様式レジストリ）が整備されている。
- 従来の院外心肺停止患者の解析は、病院到着までの心拍再開（ROSC）、一か月後の生存、社会復帰の有無をロジスティック回帰などで検討していた。
- しかし、救命士が患者に接触した時の心電図波形が、AEDが作動するVF/VT、つまり助かり易い心電図波形か否かは定かでない。
- 本報告ではJMPで接触時のVF/VTの存在率を求め、心臓突然死の基礎的背景を求める。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

ウツタイン様式レジストリとは

- 総務省消防庁が主体となって集計している、病院の外で発生した心肺停止に関するデータベース。
- 2005年～2012年で92万件近くが登録されている。
- シンガポール、台湾、韓国、日本に同様のものがあるが日本のものは高齢者が多い。

ウツタイン様式レジストリの内容

年	口頭指示あり	覚知
都道府県コード	波形種別	現着
発生年月日	除細動	接触
性別	二相性／単相性	CPR開始
年齢	初回除細動実施時刻	病院収容
救急救命士乗車	除細動施行回数	心原性／非心原性
医師の乗車	実施者：救急救命士	心原性の種別
医師の2次救命処置	実施者：救急隊員	非心原性の種別
目撃	実施者：消防隊員	心拍再開
目撃時刻	実施者：その他	初回心拍再開時刻
バイスタンダー種別	気道確保	1ヶ月予後回答
バイスタンダーCPR有無	特定行為器具使用	1ヶ月生存
心臓マッサージ	特定行為器具種別	脳機能カテゴリー
人工呼吸	静脈路確保	全身機能カテゴリー
市民等による除細動	薬剤投与	
確定／推定／不明	薬剤投与時刻	
CPR開始時刻	薬剤施行回数	

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

用語定義

- バイスタンダー　救急現場に居合わせた人
- CPR　心肺蘇生法
- BCPR　バイスタンダーによる心肺蘇生法
- AED　自動対外式除細動器
- VF　心室細動
- VT　心室頻拍（VFVT：社会復帰率高　AED作動）
- PEA　無脈性電気的活動
- Asys　心静止（PEA/Asys：社会復帰率低　AED作動不可）
- 脳機能カテゴリ　1-4：生存、5：死亡
- 　1-2：社会復帰
- 　3-4：高度障害、寝たきりなど

対象

- ウツタイン様式データ（2005-12年925288人）中、15-89歳、目撃あり、心原性疾患を対象とした。
- 非心原性疾患、市民によるAED使用は除外
- 119番通報である覚知から患者接触時間はその時間分布の99%を占める26分までとした。
- 解析対象は100388人となった。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

解析対象の内訳

❖ 2005-2012のデータを対象 (n=925288).

❖ 包含基準
- ✓ 性別が明確、年齢18-89
- ✓ 目撃あり、バイスタンダー種別が明確
- ✓ 心原性心停止、心電図波形とBCPR実施の有無に欠損値無

131301人を包含

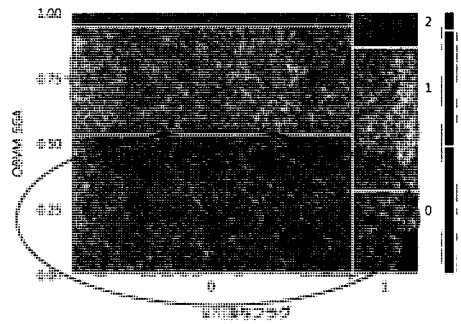
❖ 除外基準
- ✓ 薬剤投与回数不明確 28517人
- ✓ 一月後脳機能カテゴリ不明確 113人
- ✓ 気道確保の器具不明確 32人
- ✓ 覚知接触、覚知病着、覚知薬剤が 99percentile 以上の時間的異常値 1697人
- ✓ 救急隊到着前の心拍再開 232人
- ✓ 覚知接触薬剤病院着の順番が異常73人
- ✓ 初回の心拍再開後に薬剤投与 299人

30963人を除外
100338 人を 対象

薬剤投与と気道確保の器具の関係

薬剤投与フラグとOBVM-SGAの分割表に対する分析

モザイク図



分割表

薬剤投与フラグ	度数	OBVM-SGA 0	1	2	
	0	42768	34062	4130	80960
	1	6115	10667	2596	19378
		48883	44729	6726	100338

BVM-SGA
気道を確保する器具の種類
0:バッグバルブマスク
1:声門上気道デバイス
2:挿管チューブ

薬剤投与
0 投与せず
1 エピネフリン投与

本報告では、基本的手技であるBVM使用のみの48883人を対象とする

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

方法

- 救急隊接触時の心電図波形は、AEDが作動するVT/VFおよびAEDが作動しないPEA/Asysとした。
- 目的変数：覚知-接触時間（秒）
- 独立変数：年齢、性別、BCPRの有無、BCPR実施者の種類等を用いた。
- 年齢は15-49, 50-69, 70-79, 80-89の４段階に分類
- 生存期間分析、Coxの比例ハザードモデルでVF/VTをエンドポイント、PEA/ASYSを打ち切りとし、覚知-接触時間と各種変数の関係を求める。

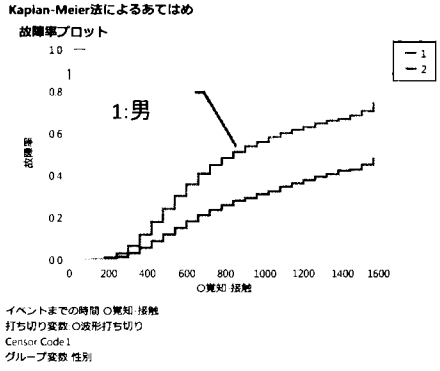
研究-1

- 比例ハザード性の検討のために、各変数のカプランマイヤー曲線を求めた。

- 補足
  - グラフ上Y軸の「故障率」の標記は、累積VF/VT存在率、つまり一種の助かり易さを意味する。
  - 覚知-接触時間＝119番入電から現地で救急隊が患者に接触するまでの時間（秒）を意味する。

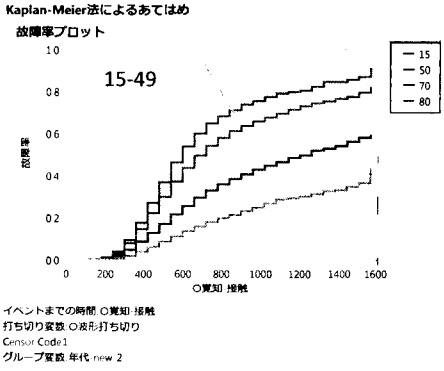
医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

性別



1:男　2:女

年代

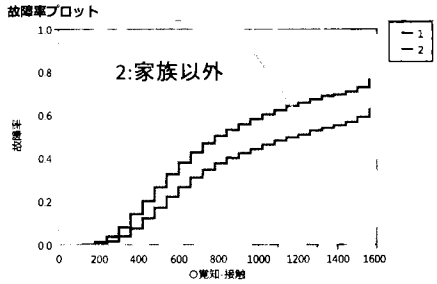


上から15-49, 50-69, 70-79, 80-89歳

バイスタンダーの種類　　昼夜

Kaplan-Meier法によるあてはめ

故障率プロット



イベントまでの時間:○覚知-接触
打ち切り変数:○波形打ち切り
Censor Code:1
グループ変数:家族か否か

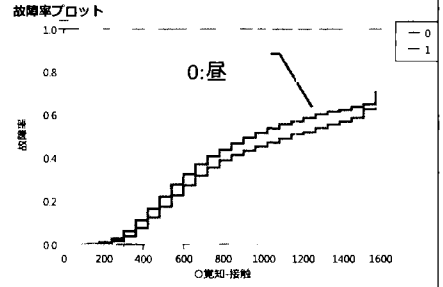
2:家族外

（友人、同僚、通行人、その他等）

1:家族

Kaplan-Meier法によるあてはめ

故障率プロット



イベントまでの時間:○覚知-接触
打ち切り変数:○波形打ち切り
Censor Code:1
グループ変数:昼夜

0:昼　6時-18時

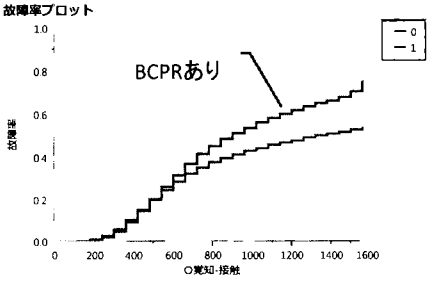
1:夜　18時-6時

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

BCPRの有無

Kaplan-Meier法によるあてはめ

故障率プロット



イベントまでの時間:○覚知-接触
打ち切り変数:○波形打ち切り
Censor Code:1
グループ変数:△BCPRあり／なし

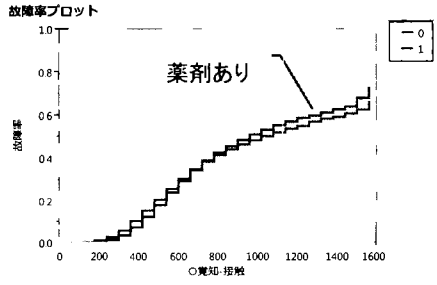
1:BCPRあり

0:BCPRなし

薬剤投与の有無

Kaplan-Meier法によるあてはめ

故障率プロット



イベントまでの時間:○覚知-接触
打ち切り変数:○波形打ち切り
Censor Code:1
グループ変数:薬剤投与フラグ

1:薬剤投与あり

0:薬剤投与無

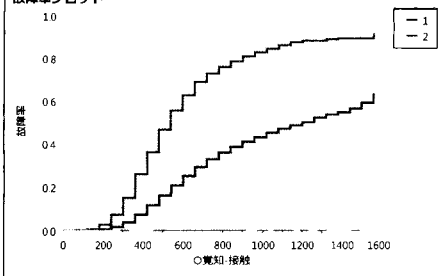
覚知接触時間との交絡が存在

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

心拍再開

Kaplan-Meier法によるあてはめ

故障率プロット



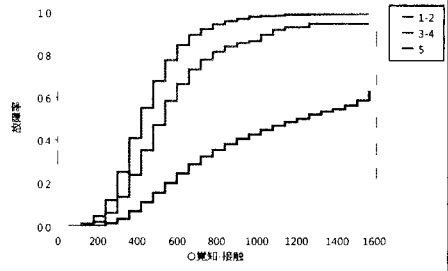
イベントまでの時間 ○覚知-接触
打ち切り変数 ○波形打ち切り
Censor Code1
グループ変数 ○心拍再開

1:心拍再開あり
0:心拍再開なし

社会復帰

Kaplan-Meier法によるあてはめ

故障率プロット



イベントまでの時間 ○覚知-接触
打ち切り変数 ○波形打ち切り
Censor Code1
グループ変数 OCPC 再分類

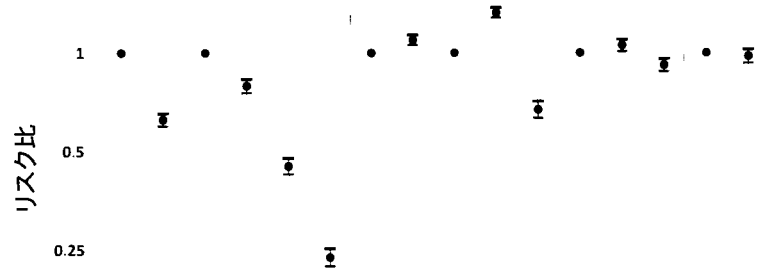
1-2:社会復帰
3-4:社会復帰不可
5:死亡

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

研究-2

- Coxの比例ハザードモデルで、VF/VTの存在率に影響を与える要因を検討した。
- BCPRの有無は覚知～接触時間で変化し、薬剤投与は県で投与タイミングが異なるが、その影響を見るため独立変数に含めた

VF/VTの存在の有無のリスク比と95%CI



	性別		年齢				BCPR		通報者		時刻			薬剤	
	男	女	15-49	50-69	70-79	80-89	BCPR無	BCPR有	家族	友人同僚等	0-6	6-12	12-18	18-24	薬剤無	薬剤有
－上側95%		0.65		0.83	0.48	0.25		1.14		1.38	0.71		1.10	0.96		1.02
●リスク比	1	0.62	1	0.79	0.45	0.24	1.00	1.10	1	1.33	0.67	1.00	1.05	0.92	1.00	0.98
－下側95%		0.60		0.75	0.43	0.22		1.06		1.28	0.63		1.01	0.87		0.93

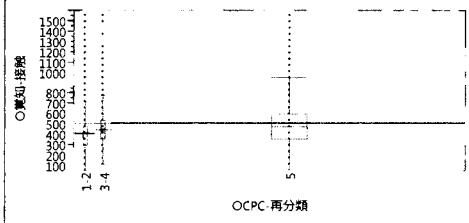
医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

研究-3

- 覚知接触時間を調整したあとでの、VF/VT存在率に影響を及ぼす変数は明らかになった。
- そこで、ひと月後の社会復帰と社会復帰不可で覚知接触時間がどうなるかを見た。

覚知(119入電)～患者接触

○CPC-再分類による○覚知-接触の一元配置分析



一元配置の分散分析

分散分析

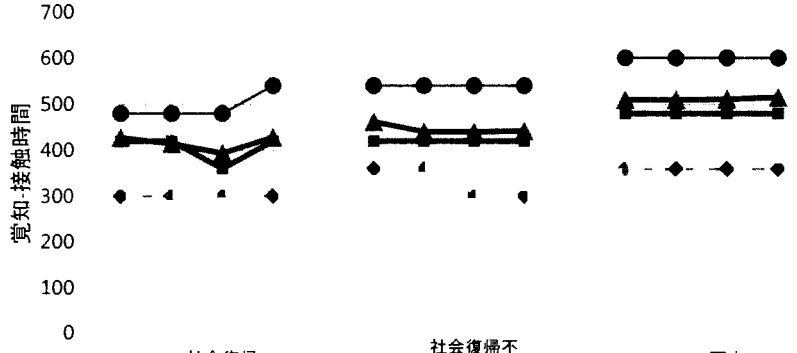
要因	自由度	平方和	平均平方	F値	p値(Prob>F)
○CPC-再分類	2	63374137.1	31687069	743.7379	<.0001
誤差	100335	4274788386	42605.157		
全体(修正済み)	100337	4338162523			

	25%値	中央値	平均	75%値	最大値
社会復帰	300	420	419.2	480	1560
社会復帰不可	360	420	454.6	540	1560
死亡	360	480	519.7	600	1560

- 社会復帰、社会復帰不可、死亡の3群で覚知-患者接触の時間に有意差がある

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

社会復帰、社会復帰不可、死亡と覚知接触時間



	社会復帰				社会復帰不可				死亡			
	15-49	50-69	70-79	80-89	15-49	50-69	70-79	80-89	15-49	50-69	70-79	80-89
75%値	480	480	480	540	540	540	540	540	600	600	600	600
中央値	420	420	360	420	420	420	420	420	480	480	480	480
平均値	427.2	414.6	393.5	428.3	462.1	440.4	440.1	442.0	510.0	509.7	511.1	515.0
25%値	300	300	300	300	360	360	300	300	360	360	360	360

まとめ-1

- 救急隊接触時のVF/VT存在率は年代と共に有意に低下し、BCPRの実施、家族以外の発見、午後の発症で有意に上昇した。
- VF/VTの存在率が低い者にエピネフリンを投与する傾向は見られなかった。
- 社会復帰は復帰不可に比べて覚知～接触時間が短い傾向にあった

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

まとめ-2

- 家族以外の発見、BCPRの実施、午後の発症は傷病者が発見されやすい環境である。
- 今後は、心肺停止時に発見されやすい環境、的確なBCPRの実施、覚知接触時間短縮のための工夫、などで院外心肺停止の社会復帰が増える可能性がある。

統計解析

東京都23区の公立図書館の比較評価

－統計とDEAの共生－

新村秀一
成蹊大学経済学部

Comparison of public libraries of 23 wards of Tokyo
- Collaboration of DEA and Statistics –

Shuichi Shinmura
Department of Economics, Seikei Univ.

要旨

統計と数理計画法を研究テーマとする筆者にとって、日本は住みにくい環境にある。分析したい対象がデータで表されておれば統計分析が、データがない場合は対象を数式で記述し数理計画法で問題解決が図れる。そこで、28歳のときにSASを、32歳のときにシカゴ大学で開発された数理計画法ソフトのLINDOを日本に紹介することにした。情報処理企業において世界最高水準の統計ソフトと数理計画法ソフトを習得し、その成果を普及することが自分にとっても日本の社会にとっても最善と考えた。しかし、私自身の中でもこの2つは容易に共生しなかった。

縁あって48歳のときに成蹊大学に移った。大学卒業後に大阪成人病センターで、心電図の自動解析システムのプロジェクトに参加し、当時脚光を浴びていた「計量診断学」を勉強し、異常心電図所見と正常所見の診断論理を判別分析でアプローチした。しかし、4年間の研究成果はプロジェクトリーダーの野村医師の開発した経験的な「枝分かれ論理」にまったく歯が立たなかった。「科学的な判別理論がなぜ？」という挫折体験が社会人の出発になった。そこで、大学での研究テーマを1998年から「学習標本での誤分類数最小化（MNM）基準による線形判別関数」を整数計画法で行うことにした。統計分野からの研究者には、「学習標本で誤分類数を最小化すれば検証標本でoverestimateすることも分かっていないのか」といわれ、数理計画法の研究者から「時間のかかる数理計画法の定式化は問題である」といわれながら、基本的な研究は2010年に研究を完成させて成果を出版した。多くのユーザーが成果に驚くと思いきや誰も驚かない。そこで「線形分離可能なデータの判別」に焦点を合わせて研究を行い、2013年に一応の成果を得たが日本ではあまり理解されない。そこで2014年は英語のフルペーパーを5本程度出すことを目標に定めて、一応の成果を得た。

次は数理計画法の研究テーマの中で、企業の経営効率性分析手法であるDEAを統計と共生することを考えている。数理計画法を用いて入力データと出力データを重回帰でアプローチするのでなく、総合化された入力と出力の比の情報を分析する手法と考えればよいであろう。DEAの研究論文の多くは、統計手法では1）回帰係数などの重みがデータ全体で共通である、2)複数の出力（目的変数）が扱えない（正準相関分析では可能）などが、DEA法が統計手法に比べて優位である、という主張が見られる。しかしこれらの主張は瑣末である。

本研究では、DEA法の結果を統計分析することで、数年前からマスコミで取り上げられた公立図書館の革命的変革を紹介する。本研究はマスコミが取り上げる1年ほど前から行っており、Big

Dataに隠れて注目されなかったのは非常に残念である。

1. はじめに

Data Envelopment Analysis(DEA)は，評価に可視化という新しい視点を持ちこんだ．入力をxとして出力yが評価を表す場合，回帰分析でy＝a*x+cという単回帰式で評価項目yがxで予測できる．この場合xは，時間的にyより先行していて，制御しやすいという条件を満たすことで実用上意味を持ってくる．欠点としてはn個の評価対象から共通の回帰係数aと定数項cを求めている点である．このとき出力yが大きいほど良いと仮定した場合，誤差eが大きくて正になるものが評価されるべきであるが，現実はyの値が大きなもの注目しがちな傾向がある．例えば，企業における事業部評価でも，売り上げや利益規模の大きな中核事業が注目され，たとえ採算性が良くても規模の小さい事業部は評価されないことが多い．これに対してDEAは，入力と出力の比を効率値y/xとしてとらえることを提案した．その上で個々のDMU_i(評価対象)に最適な重みを与えて，DEA効率値b_i*y_i/a_i*x_iを他のDMU_j(j=1,…,n)の効率値[1]を1以下にするという制約のもとで最大化することを提案した．DEAの基本的なこの手法は，米国テキサス大学のCharnesとCooper両教授とRhodesによって開発されたのでCCRモデルと呼ばれている[4][9]

$$MAX= b_i*y_i/a_i*x_i ;$$

$$b_i*y_j/a_i*x_j \leqq 1 ; \quad j=1,\cdots,n \qquad (1)$$

入力と出力が複数ある場合，入出力と重みをベクトルに置き換えて，DEA効率値を${}^tb_i*y_i/{}^ta_i*x_i$と定義すれば式(2)で一般化される．これによって重回帰分析で扱えない複数の出力変数も分析できる．

$$MAX= {}^tb_i*y_i/{}^ta_i*x_i ; \qquad (2)$$

しかし，このモデルは非線形計画法になるため，これまでは計算時間がかかり大域的探索が必要になる．そこで式(3)のように変形して，線形計画法で解くことで非線形計画法の問題が解消できる．しかし、近年急速に数理計画法ソフトの能力が向上し現時点でも研究論文で10年一日のように行われているこの定式化は意味がなくなってきている。

$$MAX= {}^tb_i *y_i ;$$

$${}^ta_i *x_i=1;$$

$${}^tb_i *y_j \leqq {}^ta_i *x_j ; \quad j=1,\cdots,n \qquad (3)$$

CCRモデルを用いる最大の利点は，評価の可視化と公平性が実現できる点である．すなわち評価対象自身に最適な重みを求めているが,その結果DEA効率値が1になる場合と,ならない場合がる．従来の企業における評価法は，上司や専門家の経験や知識に負うところが大きい．そして，その基準が分かりにくく不明であることが多く，評価が良くない場合には評価対象にとって与えられた評価が納得しにくかった．しかしDEAでは，評価対象自身に最適な重みを求めてなおかつ非効率であれば，その重みで効率値が1になる他の評価対象がいることになる．その場合，その評価対象を参照集合(手本)として改善点を考えることができる．これが重回帰分析のように共通の重みであったり，他の評価対象の重みであったり，評価基準があいまいであったりしない点が，評価の可視化や公平性を考える上で重要になる．

一方では，CCRモデルは各評価対象に一番有利な評価を行うため，入出力の変数が増えてくると

[1] DEAの目的関数の値をDEA効率値，クロス効率値(制約式)で計算されるものを効率値と区別する．

手本が増える問題がある．企業で普及を考える場合，たくさん出てくる手本の中で一つの評価対象を手本にして問題点(改善点)を検討し，必要であれば別の手本で追加検討する方が普及しやすい．CCR モデルの欠点は，手本の中で優先順位がつけられない点である．そこで式(4)の Inverted CCR モデルの利用が考えられる．DEA 効率値に代わって DEA 非効率値($^tb_i*y_i/^ta_i*x_i$)を考え，この重みを用いて他の評価対象が 1 以上になるという制約で最小化する重みを求める．このモデルの有用性は，DEA 非効率値が 1 になる非効率な手本に注目することではなく，CCR モデルで手本になった評価対象の中で DEA 非効率値が最大の評価対象を最初の改善目標にすることを提案する．

$$MIN= {}^tb_i*y_i ;$$
$$^ta_i*x_i=1;$$
$$^tb_i*y_j \geqq {}^ta_i*x_j ; \quad j=1,\cdots,n \qquad \text{式(4)}$$

以上の利点を正しく紹介し，企業へ DEA を経営効率性の改善法として普及するために以下の点を提案する．

・DEA の有効性を示す分かりやすい成功事例として，東京都の公立図書館の 1986 年と 2011 年を比較し，25 年間に目覚ましい図書館業務の改善が行われたことを示す．

・企業に DEA を普及するためには，最初の段階では CCR と Inverted CCR という基本モデルに限定する必要がある．最新の研究成果までを普及の初期段階で行うことは，多くの企業人の理解を得ることが難しく普及を困難にする．

・DEA は数理計画法で定式化され，多くのモデルが研究されている．しかし，普及のために数理計画法の理解を前提とせず，与えられた評価対象のデータ F (Factor) と，最適化で得られた重み W と，そこから計算されたクロス効率値[2]C と DEA 効率値 SCORE といったデータで説明した方が，統計分析の知識がある多くの企業人の理解が得やすい．

・普及のために，Excel 上に評価対象のデータ F を与えれば，モデルのサイズに影響を受けない CCR と Inverted CCR モデルが簡単に実行できる汎用モデル[3]を開発した [7].

・評価対象データ F の各変数の最大値を 1 以上 10 未満になるように単位を変換することで，数値計算上の問題の回避と重みの解釈が容易になる．

・1 入力 2 出力あるいは 2 入力 1 出力モデルの場合，入力と出力の 2 個の比を作り散布図を描くことで，効率的 DEA フロンティアとそれに包み込まれる非効率な評価対象の改善目標がわかる．ただし入出力変数の和が 4 個以上になると散布図を描くことはできないので，クロス効率値から求めた DEA クラスターで対応することを提案した[3][8].

・DEA は，これまでの企業における評価で単に規模が小さいことで注目されなかった評価対象であっても，DEA 効率値が 1 であれば手本であることを示してくれる．しかし，変数が多くなっていくと手本や DEA クラスターが増えていく傾向がある．評価対象全体でまず改善策を考える場合は，CCR モデルで効率的であり，Inverted CCR モデルで最大の非効率値 (逆 SCORE) をもつ評価対象を共通の改善目標と考えた方がよい．それがうまくいった後，次の改善策を考えるべきである．

・改善方法を考える場合，評価対象のデータ F を用いて，Excel で簡単に計算できる「1 入力固定改善法」を提案する．これによって経済学の学生に就活希望の業種の企業を 20 社ほど集めさせて分析させたところ、データの入力に 1 時間、分析に 2 時間ほどで、社会経験のない学生でも容易に有益な知識が得られることが分かった。しかし、テーマの選定と、分析結果をレポートにするのが多

---

[2] CCR モデルで求まる重みが無限にある場合でも，効率値が 1 になるものは影響を受けず，非効率な値だけが影響を受ける．本研究では効率値が 1 になるものだけに注目し議論を行う．

[3] 統計ソフトが普及したのは，各統計手法がデータの変更に影響を受けない点である．数理計画法の各種問題でデータの影響を受けないモデルを汎用モデルと呼ぶことにする．

くの学生のネックであることが分かった。
・専門用語としての「DMU(意思決定主体)」と「参照集合」に代わって，柔らかい印象を与える「評価対象」と「手本」に置き換えて普及した方がよいと考える．
・そして、これらの出力結果を統計分析することで、より多くの人に理解してもらえることを示す。また 1 入力と 1 出力の比がもつ情報は元のデータにない情報を持っているような官職を得ている。すなわち p 入力 q 出力のデータがあれば、p＊q 個の比を作り、それを統計分析することで元データで得られない結果が得られるのではないかと考えている。

2. 企業への DEA 普及の提案

2.1 成功事例の紹介

企業へ広く DEA を普及するには，成功事例の紹介が重要である．その点で，1986 年と 2011 年の東京都の公立図書館の事例は最適である．1986 年時点では，床面積，職員数，貸出数を用いた 2 入力 1 出力モデルで，人口の多い世田谷区と杉並区が手本となった．そして，千代田区は，住民への図書の貸し出し需要が少なく DEA 効率値は 0.19 と最低であった．これまでの研究でも，このような小さな値を持つ評価対象がある分析対象は少ない．また世田谷と他の公立図書館を「1 入力固定改善法」で比較すると，(手本の杉並を含む) 他の 22 区の公立図書館の床面積が過大であることが分かる．この場合，一番簡単な改善案は余分な床面積を貸会議室や他の文化事業などへの転用などが考えられる．しかし，2011 年時点では公立図書館の多くは，予算以上に図書館業務の拡大と改善に成功した．また 1986 年で最も非効率であった人口の一番少ない千代田区が，2011 年では効率的になった．この点を，主成分分析のスコアプロット上で効率的フロンティアに対応する曲線を描くことで，1966 年から 2011 年に効率的フロンティアが拡大したことを図で示す．

2.2 汎用モデルを用いた DEA の説明

(1) LINGO による汎用モデル

用いる DEA の手法は，付録の LINGO[2][6][7]で作成した CCR と Inverted CCR モデルである．数理計画法モデルの分析は，データのスケーリングが重要である．例えば数理計画法ソフトが 10^{-8} 以下を 0 と判定している場合，データの最大値と最小値の比が 10^{8} 以上であれば，計算過程において最大値で割ると 0 に判定されるものが出てきて数値計算上の問題が生じる．そこで DEA で分析するデータは，各変数を 10^{n} で割り最大値を 1 以上 10 未満に正規化することを提案する．これで数値計算上のトラブルが回避でき，さらに単位が明らかで重みの比較が容易になる．付録で示すが，DEA 法は式(2)で表される分数計画法を式(3)の線形計画法に変換しているため，Inverted CCR モデルで入力の重みが局所解の 0 を求めると，非効率値を計算する場合に分母が 0 になり問題が生じる．これを回避し LP で計算する方法を示す[4].

(2)　2 入力 1 出力モデルで CCR モデルの説明

図 1 は，1986 年の 23 公立図書館で，職員数（F 列）と床面積（G 列）を入力とし，貸出数(H 列)を出力とする 2 入力 1 出力モデルである．1986 年の評価対象の SN を 31 から 53 で，2011 年は 1 から 23 で区別する．評価データ F をセル範囲名[5]F(F25: H47)に与える．ただし各変数は最大値が 1

---

[4] 逆 CCR モデルでクロス効率値の計算を含めて非線形計画法モデルとして大域的探索を行えば問題が生じないが計算時間がかかる．

[5] LINGO は Excel のセル範囲名 F を「@OLE()=F;」で入力「F=@OLE();」で出力でき

以上 10 未満になるように変換してある．このデータを入力し CCR モデルを実行すれば，DEA 効率値がセル範囲名 SCORE（J25:J47）に，重みがセル範囲名 W（L25:N47）に，クロス効率値[8]がセル範囲名 C(S25:AO47)に出力される．データを基準化したことで重みの解釈がしやすくなる．床面積の大きい中央区，港区，新宿区，杉並区は床面積の重み(W2)を 0 にし，世田谷区は職員数が多いので重み(W1) を 0 にすることが DEA 効率値を高めるために有効である．重みの詳細な分析は，今後の課題とする．この千代田区の重みを評価データ F に適用し，S 列のセル範囲 S25:S47 に効率値を出力する．すなわちセル S25 は千代田区の重み(3.29, 0.64, 1.79)で計算した千代田区の効率値で 0.19 (= 3.29*0.26 +0.64 * 0.22+1.79*0.11)と非効率になる．千代田区が 1 にならないのは，千代田区の重みで計算した世田谷区(セル S36)と杉並区(セル S39)の効率値が 1 になるためである．すなわち，千代田区はこの 2 区を目標にして改善を図ればよい．DEA 以前であれば，入出力の比を比べて他の区より明らかに劣っている個々の比率が分かったとしても，それを区民人口の少なさなどに原因を帰着させて終わりになることが多かった．それが世田谷区と杉並区が手本であることが分かれば，世田谷に比べて職員数と貸出数が少なく，床面積が大きいことが簡単に分かる．実際には千代田区は，DEA を利用しないで 2011 年には目覚ましい改善を達成した．しかし，改善を考える際に，DEA の分析結果が事前に分かれば，試行錯誤の無駄が省ける．同様に中央区から江戸川区の重みを適用し，T 列から AO 列にクロス効率値を出力する．この対角要素の効率値が，J 列の DEA 効率値(セル範囲名 SCORE)である．DEA 効率値から世田谷区(セル J36)と杉並区(セル J36)の 2 区だけが手本になり，クロス効率値ではこの 2 区に対応する 36 行と 39 行の効率値だけが 1 になる．クロス効率値の S 列から AO 列の 23 個の列ベクトルで，この 2 区の効率値が 1 になるパターンは C1（千代田区の重みで計算した効率値ベクトル）と C12（世田谷区の重みで計算した効率値ベクトル）と C15（杉並区の重みで計算した効率値ベクトル）の 3 個ある．C2 は C15, C10 と C23 は C1 と同じ DEA クラスターになる．23 区の公立図書館をこの 3 個のパターンに分けて，表 1 のような DEA クラスターに分類できる．

	A	B	F	G	H	J	K	L	M	N	S	T	AB	AD	AG	AO
1	SN	区	職員数	床面積	貸出数	SCOR	逆SC	W1	W2	W3	C1	C2	C10	C12	C15	C23
25	31	千代田区	0.26	0.22	0.11	0.19	1	3.29	0.64	1.79	0.19	0.18	0.19	0.12	0.18	0.19
26	32	中央区	0.30	0.46	0.31	0.47	1.46	3.33	0	1.49	0.44	0.47	0.44	0.18	0.47	0.44
27	33	港区	0.69	1.14	0.76	0.49	1.42	1.45	0	0.65	0.45	0.49	0.45	0.18	0.49	0.45
28	34	新宿区	0.96	1.11	0.93	0.43	1.79	1.04	0	0.47	0.43	0.43	0.43	0.22	0.43	0.43
29	35	文京区	1.14	1.01	1.44	0.59	3.05	0.75	0.15	0.41	0.59	0.57	0.59	0.38	0.57	0.59
30	36	台東区	0.51	0.39	0.54	0.5	2.63	1.71	0.33	0.93	0.5	0.48	0.5	0.37	0.48	0.5
31	37	墨田区	0.61	0.54	0.84	0.64	3.3	1.4	0.27	0.76	0.64	0.62	0.64	0.41	0.62	0.64
32	38	江東区	0.77	0.62	1.10	0.67	3.53	1.12	0.22	0.61	0.67	0.64	0.67	0.47	0.64	0.67
33	39	品川区	1.27	0.93	1.16	0.44	2.26	0.69	0.13	0.38	0.44	0.41	0.44	0.33	0.41	0.44
34	40	目黒区	0.84	0.51	1.56	0.91	4.59	1.07	0.21	0.58	0.91	0.83	0.91	0.82	0.83	0.91
35	41	大田区	2.42	1.97	3.06	0.59	3.12	0.36	0.07	0.19	0.59	0.57	0.59	0.41	0.57	0.59
36	42	世田谷区	2.02	1.09	4.10	1	5.01	0	0.92	0.24	1	0.91	1	1	0.91	1
37	43	渋谷区	0.74	0.75	0.54	0.33	1.53	1.13	0.22	0.61	0.33	0.33	0.33	0.19	0.33	0.33
38	44	中野区	0.92	0.71	1.35	0.69	3.61	0.95	0.18	0.52	0.69	0.65	0.69	0.51	0.65	0.69
39	45	杉並区	1.03	1.15	2.30	1	4.28	0.97	0	0.43	1	1	1	0.53	1	1
40	46	豊島区	0.68	0.70	0.98	0.65	2.97	1.22	0.24	0.67	0.65	0.64	0.65	0.37	0.64	0.65
41	47	北区	0.96	0.78	1.35	0.66	3.47	0.9	0.17	0.49	0.66	0.63	0.66	0.46	0.63	0.66
42	48	荒川区	0.78	0.55	0.85	0.52	2.68	1.13	0.22	0.61	0.52	0.49	0.52	0.41	0.49	0.52
43	49	板橋区	1.18	1.09	1.71	0.67	3.36	0.72	0.14	0.39	0.67	0.65	0.67	0.42	0.65	0.67
44	50	練馬区	1.07	1.09	1.90	0.81	3.74	0.78	0.15	0.43	0.81	0.8	0.81	0.47	0.8	0.81
45	51	足立区	1.20	1.07	1.91	0.74	3.81	0.71	0.14	0.39	0.74	0.71	0.74	0.47	0.71	0.74
46	52	葛飾区	1.01	1.06	1.07	0.48	2.16	0.82	0.16	0.45	0.48	0.47	0.48	0.27	0.47	0.48
47	53	江戸川区	0.74	0.65	1.22	0.77	4.02	1.15	0.22	0.63	0.77	0.74	0.77	0.5	0.74	0.77

図1　汎用モデルの入力と出力結果(クロス効率値は一部のみ表示)

る．そして，LINGO 内部では F で配列計算に利用できる．

5

C15(AG列)は，杉並の重みで計算した23区の効率値であり，杉並区だけが1になる．このようなパターンになるのは杉並区を含む4区(中央(C2)，港(C3)，新宿(C4)，杉並(C15))であり，杉並区を目標に改善すればよい．杉並区を改善目標にすることは，杉並以外の3区は自分に最適な重みでなく，杉並の重み（構成比）を参考にして問題点を発見することを意味する．C1(S列)は千代田区の重みで効率値を計算し，世田谷区と杉並区が1になる．このパターンを持つのは，世田谷区と杉並区を含まない18区である．C12(AD列)は世田谷区の重みで計算し，世田谷(C12)だけが1になり構成員も世田谷だけである．DEAクラスターの利点は，入出力が4変数以上でも次の散布図と異なり対応できる点である．

表1　3個のDEAクラスター

DEA クラスター	手本	数	構成員（接頭語 C を省く）
C15	杉並	4	2-4,15
C1	世田谷，杉並	18	1,5-11,13,14,16-23
C12	世田谷	1	12

図2は，この2入力1出力モデルで，貸出数/床面積と貸出数/職員数の比を求めて散布図を描いた．原点と杉並と杉並からY軸に引いた水平線で作られる三角形がほぼDEAクラスターC15 に対応する．原点と杉並と世田谷で作られる三角形の領域がDEAクラスターC1 に対応する．原点と世田谷と世田谷からX軸に下ろした垂直線で作られる三角形の領域がDEAクラスターC12に対応する．この(X,Y)=(0,2.5)から杉並への線分と，杉並から世田谷への線分と，世田谷からX軸への垂直線で作られる折れ線をDEA効率的フロンティアと呼ぶ．全ての評価対象はこの凸体に内包される．「非効率な新宿の改善目標は，原点と新宿を結ぶ直線と世田谷と杉並を結ぶ効率的フロンティアの線分の交点が改善目標(DEA効率値1)である」という説明が行われているが厳密には正しくない．同じことが千代田区でもいえる．原点と千代田区を結ぶ直線と世田谷からX軸に下ろした垂線の交点が改善目標ではない．表1に示すように，新宿は中央区と港区と同じDEAクラスターC15に，千代田区(C1)は DEAクラスターC1に属している．また実態のない理想点を改善目標に選ぶことは，理想点の最適な重み(構成比)を参考にして問題点を考えることになるが，実際の手本との比較で問題を見つける方が現実的で説得力が出てくる．

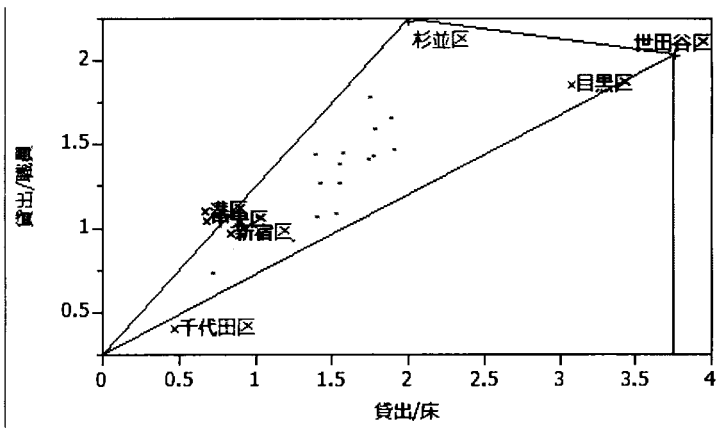


図2　効率的フロンティアと非効率な評価対象の改善目標の関係

この散布図による説明は入力と出力の1変数毎の組み合わせであり，DEAの効率値は重みで総合化された多入力と多出力の比で定義しているので，厳密な説明には利用できない．例えば，中央区，港区，新宿区の床面積の重み(W2)は図1から0であることが最適であり，この3区にとってX軸の値の違いは意味がない．世田谷や杉並も個々に異なった重みを用いているので，この図による説明は誤解を生じるので利用に際して注意がいる．

(3) Inverted CCRモデルの利点

汎用モデルでCCRモデルの分析後，Inverted CCRモデルが実行される．表2のSCOREはCCRモデルのDEA効率値で，それ以降がInverted CCRモデルの非効率値（逆SCORE）と重みとクロス効率値の一部である．クロス効率値は，23区の重み全てで千代田区が非効率値1になった．非効率値が最大なのは世田谷区の5.01であり，次は手本でない目黒区の4.59であり，手本の杉並区は4.28で目黒区より小さい．2つの異なった基準で，世田谷が 1986年時点の2入力1出力モデルで，他の図書館が共通して改善目標にすべきことが分かる．この場合，表1の杉並区を手本とした4区の扱いは後で検討する．

表2　Inverted CCRモデル

SN	区	SCORE	逆 SCORE	W1	W2	W3	C1	C10	C11	C12	C15
31	千代田区	0.19	1	3.85	0	9.49	1	1	1	1	1
40	目黒区	0.91	4.59	1.19	0	2.94	4.59	4.59	4.59	4.59	6.57
41	大田区	0.59	3.12	0.41	0	1.02	3.12	3.12	3.12	3.12	3.31
42	世田谷区	1	5.01	0.5	0	1.22	5.01	5.01	5.01	5.01	8.03
45	杉並区	1	4.28	0	0.87	1.86	5.51	5.51	5.51	5.51	4.28

2.3　1入力固定改善法

CCRモデルとInverted CCRモデルを併用して，最初の改善目標を世田谷に決めた．次に他の公立図書館の問題点を発見する方法の一つとして「1入力固定改善法」を説明する．表3の3列から5列は代表的な5区のデータである．世田谷を目標として，床面積(m^2)を固定して考える．そして世田谷以外の図書館の職員数(人)と貸出数(冊)を世田谷の構成比と同じになるよう比例計算する．例えば杉並区を考えると次のようになる．

（世田谷の）床面積:職員数:貸出数=10888 : 202:4096300

=11469/10888*(10888 : 202: 4096300)

= 11469: 213: 4314885=杉並の改善目標

表の6列から8列はこの改善目標値である．この値を達成できれば世田谷の重みで全ての区のDEA効率値は1になる．そして実際の値から改善目標値を引いたものが9列から11列になる．負であれば現在の値が世田谷の構成比に比べて少ないので改善が必要になる．一方で正の場合は，世田谷基準を上回っているので現状維持するか，少し削減し負の入力を増やすかのトレード・オフを考えることになる．この計算はExcelで簡単に計算できる．杉並区は図書館の収容力に対して職員が110人少なく，貸出数が2,015,191冊少ないことが分かる．杉並区は職員を110人増やして，貸出数を2,015,191冊と現状の2倍に増やせるか検討することになる．このような非常識な値になるのは，杉並や大田区の床面積が必要以上に広すぎるためである．そこで余分な床面積を貸会議室などに転用し，縮小均衡を図ることが現実的で容易である．例えば杉並の床面積を6000と半分にしてCCRモデルを解くと世田谷のDEA効率値が0.98になり，杉並だけが手本になる．すなわち手本である杉並であっても，世田谷に比べて床面積が過大であり改善すべき問題点が分かる[6]．

表に載せた区を含め世田谷を除くすべての22区の床面積以外の入出力値が負になる．また入力に蔵書数，出力に登録者数を加えた3入力2出力モデルで計算しても同じ結果になる．しかし，後で分析する5入力2出力モデルでは，予算と人口を入力に登録者数を出力に取り込むと手本などが大幅に異なってくる．

表3　床面積を固定した世田谷区の構成比による1入力固定改善法

[6] 本論文では，これ以降このような個別の変数の改善は議論しない．

SN	区	床面積	職員数	貸出数	床面積	職員数	貸出数	床面積	職員数	貸出数
31	千代田	2249	26	105321	2249	42	846122	0	-16	-740801
40	目黒	5077	84	1562274	5077	94	1910077	0	-10	-347803
41	大田	19716	242	3055193	19716	366	7417584	0	-124	-4362391
42	世田谷	10888	202	4096300	10888	202	4096300	0	0	0
45	杉並	11469	103	2299694	11469	213	4314885	0	-110	-2015191

「1入力固定改善法」の問題点は，図2で説明した欠点と同じく，22区の改善目標値が生産可能集合(あるいは効率的フロンティア)をはみ出すこともある点である．しかし，改善目標を現実に実現できるか否かを検討し，例え現状の生産可能集合をはみ出していても，結果として達成可能であれば問題がないと考える．計画が達成できなければ，結果責任を問えば済むことである．一方，改善目標が生産可能集合の中にあっても，各公立図書館の改善能力が低ければ改善目標はクリアできない．すなわち，生産可能集合をはみ出す可能性に注意して，「1入力固定改善法」を利用すればよい．あるいは固定した入力変数の改善目標に対する比が問題であることを示しているので，固定した入力の改善を考えた方が現実的で簡単である．

一方，この改善法の利点は次のとおりである．

・1入力に限定し，改善目標に選んだ世田谷の構成比に比例した改善目標値と現実の値との差の計算は簡単にでき，内容の理解も容易である．

・単純な比率の比較は企業でも良く行われていて，改善活動にとって多くの関係者が理解しやすいという点で重要である．

・「1入力固定改善法」は，固定する変数の違いで複数の代替案が得られ，それらを比較することで不完全であるが総合化して判断できる．多入力と多出力で適切な改善法が分かっても，多くの企業人が簡単に理解できなければ普及は難しい．

・分析に用いていない蔵書数と登録者数を加えて「1入力固定改善法」で検討しても同じ結果になるので，3入力2出力モデルのDEAの分析は省略できる．ただし分析を行うと世田谷と杉並に加えて板橋区が手本に加わる．

3. 1986年と2011年の5入力2出力モデルによる検討

3.1 単年度ごとの検討

表4(左)は1986年の23公立図書館の予算，区の人口，床面積，蔵書数，職員数を入力とし，貸出数と登録者数を出力とする5入力2出力[7]のCCRモデルとInverted CCRモデルによる分析結果である．7図書館が手本になった．Inverted CCRモデルから千代田区に加え，台東区と江戸川区が非効率な手本になった．またCCRモデルで手本のうち，Inverted CCRモデルで文京区の効率値が2.65と最大になる．

表4　1986年(左)と2011年(右)のCCRとInverted CCRモデルによる比較

SN	区	SCORE	逆SCORE	SN	区	SCORE	逆SCORE
31	千代田区	0.35	1	1	千代田区	1	1
33	港区	1	1.42	3	港区	1	1

[7] 5入力2出力を検討するのは，1986年のデータとして文献[9]に記載されているものを採用したためである．

35	文京区	1	2.65		
36	台東区	0.57	1		
40	目黒区	1	2.45		
41	大田区	0.79	1.48		
42	世田谷区	1	2.27		
45	杉並区	1	1.64		
47	北区	1	1.10		
49	板橋区	1	1.58		
53	江戸川区	0.79	1		

4	新宿区	0.67	1
5	文京区	1	1.5
7	墨田区	0.51	1
8	江東区	1	1
10	目黒区	1	2.02
11	大田区	1	1.05
12	世田谷区	0.86	1.30
13	渋谷区	0.57	1
18	荒川区	0.70	1
19	板橋区	1	1.09
20	練馬区	1	1.38
21	足立区	1	1.20
23	江戸川区	1	1.18

表 4(右)は 2011 年の CCR と Inverted CCR モデルによる分析結果である．1986 年に非効率値が 1 の千代田区と江戸川区さらに公立図書館改革の先鞭をつけた足立区を含む 10 図書館が，DEA 効率値 1 の手本になった．Inverted CCR モデルから 7 図書館が非効率な手本になり，目黒区の非効率値が 2.02 と最大になった．また江東区は両方とも 1 である．CCR モデルは改善目標になる手本を客観的に示してくれるが，変数が多いか評価対象が多様化すれば多くの手本を見つけ，改善活動が細分化されて改善目標があいまいになる．それを避けるため，Inverted CCR モデルで非効率値が最大のものを最初の改善目標と考える．

3.2 両年度の 46 図書館の検討

表 5 は 2011 年と 1986 年の 46 公立図書館の分析結果である．手本は，表 4（右）の 2011 年度単独の分析と同じ 10 図書館で DEA 効率値(SCORE)も全て同じである．すなわち 1986 年のデータにまったく影響されないことが分かる．1986 年単独で手本であった表 4（左）の 7 図書館の効率値は，表 5 の文京区(SN=35)の 0.65 から世田谷区(SN=42)の 0.89 の間にある．以上から 1986 年単独の効率的フロンティアは表 5 では非効率になり，2011 年の効率的フロンティアに図書館業務が拡大したことが分かる．また Inverted CCR モデルから 2011 年の目黒区(SN=10)の非効率値が 5.35 と一番大きい．

表 5　2011 年度と 1986 年度の 5 入力 2 出力モデルによる比較

SN	区	予算	人口	床面積	蔵書数	職員数	貸出数	登録者数	SCORE	逆 SCORE
1	千代田区	0.16	0.51	0.37	0.30	0.99	0.81	0.70	1	2.03
3	港区	1.48	2.28	1.37	0.83	0.42	2.53	2.07	1	2.57
5	文京区	1.11	2.00	1.19	1.04	0.31	3.64	1.80	1	4.62
7	墨田区	0.43	2.5	0.63	0.67	0.57	1.29	0.64	0.51	2.36
8	江東区	0.82	4.74	1.76	1.48	0.59	4.59	0.97	1	1.85
10	目黒区	0.40	2.62	0.99	1.15	0.93	4.65	2.08	1	5.35
11	大田区	1.50	6.94	2.13	1.71	0.16	4.82	1.92	1	2.65
12	世田谷区	0.71	8.53	1.83	1.99	3.06	6.68	3.11	0.86	3.50
15	杉並区	1.16	5.39	1.95	2.28	1.15	5.05	2.10	0.69	2.71
19	板橋区	1.07	5.36	1.80	1.31	0.24	3.45	2.21	1	3.01
20	練馬区	1.53	7.08	1.98	1.64	1.27	6.75	2.53	1	3.42

21	足立区	0.47	6.66	1.99	1.77	1.21	3.30	2.70	1	2.32
23	江戸川区	1.79	6.80	2.17	1.24	1.29	5.34	2.45	1	3.34
31	千代田区	0.16	0.49	0.22	0.16	0.26	0.11	0.06	0.17	1
33	港区	0.31	1.92	1.14	0.36	0.69	0.76	0.57	0.70	1.42
35	文京区	0.38	1.94	1.01	0.54	1.14	1.44	0.66	0.65	2.65
36	台東区	0.15	1.76	0.39	0.28	0.51	0.54	0.16	0.47	1
40	目黒区	0.21	2.67	0.51	0.51	0.84	1.56	0.65	0.75	2.45
41	大田区	0.75	6.60	1.97	1.26	2.42	3.06	0.98	0.59	1.48
42	世田谷区	0.59	8.08	1.09	1.15	2.02	4.10	1.91	0.89	2.27
45	杉並区	0.57	5.38	1.15	0.77	1.03	2.30	0.85	0.72	1.64
47	北区	0.16	3.66	0.78	0.53	0.96	1.35	0.37	0.73	1.10
49	板橋区	0.90	5.04	1.09	0.57	1.18	1.71	1.03	0.82	1.58
53	江戸川区	0.33	5.17	0.65	0.47	0.74	1.22	0.47	0.63	1

注:2011年の職員数は，調査票の常勤と非常勤の合計を用いた．臨時職員は0記入の区が多いので含まない．また大田区，中野区，北区，板橋区は非常勤が0になっている．千代田区はすべて外注化しているので0と表記されていたので，HPに公開されている人数を用いた．

3.3 2011年と1986年の増減比率と入出力比の検討

表6は，2011年と1986年の増減比率である．増減比率は，年2.81%で25年間毎年延びた場合に2になるので，2以上か以下かに注目する．入力で2倍以上の区は，予算が16区，人口は0，床面積は2区，蔵書数は11区，職員数は4区である．それ以上に出力の貸出数は18区，登録者数は15区と大きく図書館業務が拡大している．

予算は16区と多いが，2未満に千代田区を含む7図書館がくる．特に本研究で注目する千代田区，目黒区，大田区，世田谷区と足立区が含まれていて，これらの区は予算に比べて図書館業務を改善したことが分かる．千代田区は予算が25年間で4%，人口は3%しか伸びていないが，貸出数で7.71倍，登録者数で12.57倍と著しく伸びていて，1986年に最も非効率な状態から2011年には手本になった．あるいは1986年には，法人税などで区の財政に余裕があり放漫な予算であったともいえる．人口は6区で減少している．床面積が2倍以上は，江東区と江戸川区だけであり予算も4.17と5.4倍と増えていて，図書館サービスに力を入れたことが分かる．蔵書数は，11区が2倍以上である．職員数は4区が2倍以上で，10区が1未満と減少している．図書館業務が著しく増えているので，職員数の減少は考えられず記載の不統一のためと考える．特に大田区の職員数は242人（図1）が16人（表5の四角い枠）と226人の減少は大きい．足立区のように中央図書館だけが直轄で，分館の外部委託が考えられる．0.07という増加率は区の職員を97%減らしたことを表すと考えられる．18区の貸出数と15区の登録者数が2倍以上で，図書館の業務量は大きく増加したことを表す．千代田区と目黒区に代表される公立図書館は，この25年間に予算以上に図書館業務を拡大したと評価できよう．一方，大田区や世田谷区は2倍以下であるが，1986年時点ですでに規模が大きいので健闘していると考えるべきである．

このように個々の比率で分析できるが，統一性を欠く点である．DEA効率値と非効率値はそれを統一的に判断する基準を与えたと評価できる．

表6　2011年の1986年に対する増減比率

	予算	人口	床面積	蔵書数	職員数	貸出数	登録者数
千代田区	1.04	1.03	1.66	1.82	2.02	7.71	12.57

中央区	5.02	1.57	1.45	1.90	1.23	4.66	4.55
港区	4.76	1.19	1.20	2.29	0.61	3.33	3.61
新宿区	3.28	0.97	1.15	1.71	2.17	2.61	2.16
文京区	2.93	1.03	1.18	1.91	0.27	2.53	2.72
台東区	2.67	1.03	1.59	1.99	1.18	3.37	6.58
墨田区	2.47	1.10	1.16	1.31	0.93	1.54	1.82
江東区	4.17	1.22	2.83	3.76	0.79	4.17	1.68
品川区	2.97	1.02	1.20	1.66	0.46	2.88	1.49
目黒区	1.91	0.98	1.94	2.24	1.11	2.98	3.18
大田区	1.99	1.05	1.08	1.36	0.07	1.58	1.96
世田谷区	1.19	1.06	1.68	1.73	1.52	1.63	1.63
渋谷区	3.34	0.87	1.45	2.50	0.43	3.03	1.91
中野区	2.08	0.94	1.39	1.80	0.24	1.70	1.49
杉並区	2.04	1.00	1.70	2.96	1.15	2.19	2.48
豊島区	3.32	0.96	1.42	1.85	1.41	2.25	2.94
北区	6.51	0.91	1.78	2.37	0.66	2.80	5.20
荒川区	3.37	1.08	1.34	1.74	1.50	2.36	1.80
板橋区	1.19	1.06	1.65	2.31	0.20	2.02	2.15
練馬区	4.52	1.20	1.82	2.45	1.18	3.55	3.63
足立区	1.06	1.07	1.85	2.09	1.11	1.73	3.02
葛飾区	1.13	1.08	1.52	2.14	2.39	3.31	3.93
江戸川区	5.40	1.32	3.33	2.66	2.14	4.36	5.19

表７は，表６の増減比率を用いた入出力比の比較である．最初の５列は貸出数と５入力の比であり，次の５列は登録者数と５入力の比であり，最後の列は貸出数/登録者数でリピーター率に対応している．貸出数との比で２倍以上は２区，17区，９区，２区，15区であり，登録者数の比で２倍以上は４区，16区，９区，４区，17区であり，ほぼ同じである．予算と蔵書に対する貸出数と登録者数が２倍以上になった区は２区か４区である．人口と職員数に対する貸出数と登録者数が２倍以上になった区は15区から17区と多い．床面積に対する貸出数と登録者数が２倍以上になった区は９区と中間である．以上から，25年間で人口と職員数に対して図書館業務が著しく改善した．これは開館時間の拡大で人口がそれほど増えない中にあって利用者層の拡大を図り，図書館業務の拡大を一部外注化などで乗り切って出力を増加させたためと考える．これに対して，予算や蔵書数の増加に対し，貸出数と登録者数の伸びは小さかった．表６と表７から公立図書館ごとに当事者はさらに改善策を詳細に検討できるが，本稿の目的と外れるので省略する．

最後の列は，貸出数/登録者数の比で，千代田区を含む15区が１以下であり，リピーターが少ないようだ．台東区だけが２以上で，リピーターが多いことが分かる．千代田区にとってこの指標だけが悪い．次の経営効率化の改善目標として，台東区を調べてリピーター率を上げることが考えられる．

表７　2011年の1986年に対する入出力比

	貸出/予算	貸出/人口	貸出/床面積	貸出/蔵書	貸出/職員	登録/予算	登録/人口	登録/床面積	登録/蔵書	登録/職員	貸出/登録
千代田	7.44	7.46	4.65	4.24	3.81	12.13	12.15	7.59	6.92	6.22	0.61

中央区	0.93	2.97	3.20	2.45	(3.78)	0.91	2.90	3.13	2.40	3.69	1.02
港区	0.70	2.81	2.78	1.45	5.47	0.76	3.04	3.01	1.58	5.93	0.92
新宿区	0.80	2.70	2.27	1.52	1.20	0.66	2.24	1.88	1.27	1.00	1.20
文京区	0.86	2.46	2.15	1.32	9.32	0.93	2.65	2.30	1.42	10.01	0.93
台東区	1.27	3.27	2.12	1.70	2.86	2.47	6.39	4.13	3.32	5.58	0.51
墨田区	0.63	1.41	1.33	1.18	1.65	0.74	1.66	1.57	1.39	1.94	0.85
江東区	1.00	3.43	1.47	1.11	5.26	0.40	1.38	0.59	0.45	2.12	2.48
品川区	0.97	2.82	2.39	1.74	6.20	0.50	1.46	1.24	0.90	3.20	1.94
目黒区	1.56	3.03	1.53	1.33	2.67	1.67	3.25	1.64	1.42	2.86	0.93
大田区	0.79	1.50	1.46	1.16	(23.84)	0.98	1.86	1.81	1.45	29.63	0.80
世田谷	1.37	1.54	0.97	0.94	1.08	1.37	1.54	0.97	0.94	1.07	1.00
渋谷区	0.91	3.49	2.10	1.21	7.06	0.57	2.19	1.32	0.76	4.44	1.59
中野区	0.82	1.81	1.23	0.94	7.10	0.72	1.59	1.08	0.83	6.24	1.14
杉並区	1.07	2.19	1.29	0.74	1.92	1.21	2.47	1.46	0.84	2.16	0.88
豊島区	0.68	2.35	1.59	1.22	1.50	0.89	3.06	2.07	1.59	1.96	0.77
北区	0.43	3.07	1.58	1.18	(4.27)	0.80	5.70	2.92	2.19	7.92	0.54
荒川区	0.70	2.18	1.76	1.35	1.57	0.54	1.66	1.34	1.03	1.20	1.31
板橋区	1.70	1.90	1.22	0.87	(9.94)	1.81	2.02	1.30	0.93	10.56	0.94
練馬区	0.79	2.96	1.95	1.45	3.00	0.80	3.03	2.00	1.48	3.07	0.98
足立区	1.63	1.62	0.93	0.83	1.56	2.84	2.82	1.63	1.44	2.72	0.57
葛飾区	2.93	3.07	2.17	1.54	1.39	3.48	3.65	2.58	1.83	1.65	0.84
江戸川	0.81	3.32	1.31	1.64	2.04	0.96	3.95	1.56	1.95	2.42	0.84

3.4 1入力固定改善法

(1) 目黒区を最初の改善目標とすることの妥当性

CCR モデルで手本に選ばれた 10 区から，Inverted CCR モデルで非効率値が最大になる目黒区を最初の改善目標とすることの妥当性を以下で検討する．目黒区は予算/人口が 21 位で，貸出総数は 6 位，登録者数は 7 位であり，これまでの評価法では注目されない区である．予算/人口が少ない割に出力がある程度良くて予算と貸出数の重みだけが正で手本になった．表 8 は，DEA 効率値(SCORE)を第 1 ソートキー，非効率値を第 2 ソートキー(逆 SCORE)として降順で並べ替えた．DEA 効率値が 1 の 10 区を手本とする区の数は，目黒区は港区と江戸川区を除く 22 区の手本になった．港区の重みは人口と職員数と登録者数が正であり，目黒区の効率値が 0.83 で千代田区，港区，文京区の効率値を 1 にする．表 6 の予算の増加率 2 位の江戸川区の重みは蔵書数と貸出数と登録者数が正の重みで，予算の少ない目黒区の効率値が 0.93 で予算の多い港区と江戸川区の効率値を 1 にする．以上から少なくとも目黒区は 21 区の手本になると考えられ，港区と江戸川区の扱いを別途検討する必要がある．

一方，DEA 効率値と非効率値が 1 となる区は，改善率が大きい千代田区，予算/人口が 1 位の港区，貸出数が 7 位の江東区，予算/人口が 23 位と最も少ない足立区といった特徴をもつ 4 区である．非効率値だけが 1 になるのは，荒川区，新宿区，渋谷区，墨田区の 4 区である．前者の 4 区は何か特徴が明確で，それらを手本とする区は 3 区から 7 区と少ない．これに対し，後者に含まれる渋谷区は 20 区，墨田区は 19 区と多い．

表 8 DEA 効率値と非効率値で並べ替えた 23 区

順位	SN	区	SCORE	手本数	逆 SCORE	手本数
1	10	目黒区	1	21	2.017	
2	5	文京区	1	9	1.498	
3	20	練馬区	1	2	1.379	
4	23	江戸川区	1	6	1.180	
5	19	板橋区	1	4	1.093	
6	11	大田区	1	5	1.054	
7	1	千代田区	1	7	1	3
8	3	港区	1	6	1	7
9	8	江東区	1	2	1	5
10	21	足立区	1	1	1	6
11	6	台東区	0.952		1.580	
12	14	中野区	0.876		1.066	
13	9	品川区	0.871		1.090	
14	17	北区	0.858		1.511	
15	12	世田谷区	0.855		1.305	
16	22	葛飾区	0.817		1.203	
17	16	豊島区	0.798		1.154	
18	2	中央区	0.796		1.193	
19	18	荒川区	0.697		1	4
20	15	杉並区	0.685		1.144	
21	4	新宿区	0.669		1	11
22	13	渋谷区	0.574		1	20
23	7	墨田区	0.508		1	19

(2) 目黒区を手本にした「1入力固定改善法」

表9は目黒区を改善目標として，表8の手本である文京区，江戸川区，大田区，千代田区，港区の5区と非効率な世田谷区と墨田区の2区で「1入力固定改善法」を行った．5個の入力変数に対して5個の「1入力固定改善法」の代替案がある．これを目黒区との比の大きな入力変数の順に固定して並べ替えた．そして，比較のため出力の一つが最小の正になる代替案で比較する．

最初の5行は文京区の入力変数を予算，床面積，蔵書数，人口，職員数の順に固定して「1入力固定改善法」を行った．目黒区の各入力変数の値との比をとると，この順に小さくなる．文京区は目黒区と比較して予算の比が一番大きいので，他の入出力変数の値は全て負になる．床面積の比は予算より小さくて残りの3変数より大きいので，床面積で固定すると予算だけ正で残りは負になる．以上からこのような順に並べ替えた表の入出力の値は昇順に改善される．そして入力変数で作られる配列の対角要素は固定したことを表す0になり，対角セルの上は負に下は正の値になる構造をもつ．最初の代替案は予算を固定しているので改善目標値が一番大きくなり，表に示す出力は一番悪くなる．最後の行は職員数の比が一番小さいので，改善目標値は一番小さくなり，表に示す出力は一番良くなる．代替案の真ん中である3番目で他の区との比較することが考えられるが，出力のうちの一つが最初に正になった4行目の人口を固定した代替案で比較する．目黒区と比べて予算は8.1億円，床面積が0.44万m^2，蔵書数が17万冊多く，職員数が40人少ない．貸出数は11万冊，登録者数は2.2万人多い．文京区は目黒区に対して予算，床面積，蔵書数に余裕があるので，予算を工夫して職員数を40人増やせば，さらに図書館業務の改善が行える可能性がある．

江戸川区は，入力変数を予算，人口，床面積，職員数，蔵書数の順に固定して「1入力固定改善法」を行った．入力を順次固定していくと，予算が目黒区に比べて最大13.6億円多いのに，蔵書数が最大386万冊少ない問題がある．25年間で予算を5.4倍増やしたが，蔵書数を増やすことに気づかなかったようだ．少ない蔵書数を固定すると，全ての入出力が正になる．予算を工夫し蔵書数を増やすことができれば，目黒区に比べて床面積や職員数に余裕があり，さらに図書館業務を改善できると考えられる．

大田区は職員数を固定すると他の入力は全て正になり，貸出数は402万冊，登録者数は15.6万人多く問題がないように見える．DEAの分析では調査データの職員数を用いているが，次に回帰分析で職員数を予測し再検討する．職員数を他の6変数で変数選択を行い，予算と人口の回帰式(職員数=0.37-0.28*予算+0.19*人口)が得られた．予測値は130人になり，この値で「1入力固定改善法」を行うと参考行の結果になる．職員数を固定すると，貸出数が169万冊，登録者数が10万人少なくなる．またDEA効率値は0.69，非効率値は2.65となり非効率になる．世田谷区とともに図書館業務の規模の大きな区の未来像を模索するか，目黒区を手本に改善するかが考えられる．

千代田区は，入力変数を職員数，予算，床面積，蔵書数，人口の順に「1入力固定改善法」を行った．蔵書数を固定すると，職員数，予算，床面積，登録者数が多く，人口は1.7万人，貸出数が39万冊少ない．表7でも指摘したが登録者のリピート率を上げることが問題点として考えられる．

港区は，入力変数を予算，床面積，人口，蔵書数，職員数の順に「1入力固定改善法」を行った．人口を固定すると，予算と床面積と登録者数が2.6万人多く，蔵書数が16万冊，職員数が39人，貸出数が151万冊少ない．25年間で予算を4.76倍増やしたが，蔵書数と職員数を増やす配分が悪かったようだ．

表9　目黒区を改善目標とする「1入力固定改善法」

SN	区	予算	床面積	蔵書数	人口	職員数	貸出数	登録者数
5	文京区	0	-1.53	-2.13	-5.26	-2.26	-9.21	-3.96
5	文京区	0.63	0	-0.35	-1.18	-0.81	-1.97	-0.72
5	文京区	0.75	0.3	0	-0.38	-0.53	-0.57	-0.09
5	文京区	0.81	0.44	0.17	0	-0.4	0.11	0.22
5	文京区	0.98	0.86	0.65	1.12	0	2.09	1.1
SN	区	予算	人口	床面積	職員数	蔵書数	貸出数	登録者数
23	江戸川区	0	-4.89	-2.23	-2.85	-3.86	-15.39	-6.83
23	江戸川区	0.75	0	-0.39	-1.12	-1.73	-6.72	-2.95
23	江戸川区	0.91	1.04	0	-0.75	-1.27	-4.88	-2.12
23	江戸川区	1.23	3.15	0.79	0	-0.35	-1.14	-0.45
23	江戸川区	1.36	3.95	1.1	0.28	0	0.28	0.19
SN	区	予算	人口	床面積	蔵書数	職員数	貸出数	登録者数
11	大田区	0	-2.82	-1.54	-2.56	-3.3	-12.5	-5.83
11	大田区	0.43	0	-0.48	-1.33	-2.3	-7.49	-3.59
11	大田区	0.63	1.27	0	-0.77	-1.85	-5.24	-2.58
11	大田区	0.9	3.04	0.66	0	-1.22	-2.11	-1.18
11	大田区	1.43	6.49	1.96	1.51	0	4.02	1.56
参考	大田区	0.93	3.27	0.75	0.1	0	-1.69	-1
SN	区	職員数	予算	床面積	蔵書数	人口	貸出数	登録者数

1	千代田区	0.00	-0.26	-0.68	-0.92	-2.29	-4.14	-1.52
1	千代田区	0.61	0.00	-0.03	-0.17	-0.56	-1.09	-0.15
1	千代田区	0.64	0.01	0.00	-0.14	-0.48	-0.95	-0.09
1	千代田区	0.75	0.06	0.12	0.00	-0.17	-0.39	0.16
1	千代田区	0.81	0.09	0.18	0.08	0.00	-0.09	0.30
SN	区	予算	床面積	人口	蔵書数	職員数	貸出数	登録者数
3	港区	0	-2.27	-7.4	-3.39	-3.01	-14.63	-5.61
3	港区	0.93	0	-1.35	-0.75	-0.87	-3.91	-0.82
3	港区	1.13	0.51	0	-0.16	-0.39	-1.51	0.26
3	港区	1.19	0.65	0.38	0	-0.25	-0.85	0.55
3	港区	1.3	0.92	1.09	0.31	0	0.42	1.13
SN	区	職員数	人口	床面積	予算	蔵書数	貸出数	登録者数
12	世田谷区	0	-0.11	-1.42	-0.62	-1.79	-8.65	-3.75
12	世田谷区	0.04	0	-1.38	-0.6	-1.74	-8.44	-3.66
12	世田谷区	1.34	3.67	0	-0.04	-0.14	-1.95	-0.75
12	世田谷区	1.43	3.92	0.1	0	-0.02	-1.49	-0.55
12	世田谷区	1.45	3.98	0.12	0.01	0	-1.39	-0.5
SN	区	予算	人口	床面積	職員数	蔵書数	貸出数	登録者数
7	墨田区	0.00	-0.33	-0.44	-0.43	-0.57	-3.73	-1.61
7	墨田区	0.05	0.00	-0.31	-0.32	-0.43	-3.14	-1.35
7	墨田区	0.18	0.83	0.00	-0.02	-0.07	-1.67	-0.69
7	墨田区	0.19	0.89	0.02	0.00	-0.04	-1.56	-0.64
7	墨田区	0.20	0.98	0.06	0.03	0.00	-1.41	-0.57

手本である5区の場合，出力が正になるものがあるが，非効率な世田谷区や墨田区は出力が全て負である．世田谷区を蔵書数で固定しても，貸出数は139万冊，登録者数は5万人少ない．この場合，余裕のある職員数，人口，床面積，予算，蔵書数の順に検討して無駄を省き，蔵書数，予算の順に増やすことを検討し出力を改善することが考えられる．世田谷区は図書館運営を区の直営として区の職員を手厚く充当している．このため職員数が目黒区に対し最も非効率であり，他の入力を固定すると4人から145人多い．一方，蔵書数は2万冊から179万冊少ない．1936年に世田谷区と比べると，他の22区は床面積が2倍以上の区もあり非効率的であった．しかし，職員数を固定すると目黒区に比べて1.42万m^2，人口を固定すると1.38万m^2少なくなっているので，世田谷区は25年間で床面積が狭くなり問題を抱えていないか検討すべきである．

墨田区は蔵書数を固定すると，予算が2億円，人口が9.8万人，床面積が0.06万m^2，職員数が3名多く，貸出数が141万冊，登録者数が5.7万人少ないので目黒区を参考に大幅な改善策を検討すべきである．

4. 統計分析と資料による検討

4.1 主成分分析

図3は，5入力2出力に用いた7変数で主成分分析し，第1主成分と第2主成分上でスコアプロ

ットを描いた[1][5]．そして 1986 年と 2011 年のデータを個別に 95%正規確率楕円を描いた．大きな正規確率楕円が 2011 年である．それに含まれる左上の小さなものが 1986 年である．3 象限にある 1986 年の千代田区は，2011 年には第 2 象限の＋印の右上に移動している．移動距離から，千代田区の図書館業務の規模が小さいことが分かる．1986 年に手本であった第 1 象限の世田谷区は，右上の+に移動し業務を大きく拡大して外れ値になっている．ほぼ重なっている大田区が第 4 象限に大きく移動しているが，移動方向はほぼ 90 度異なっている．第 2 象限にある目黒区は，2011 年には第 4 象限に大きく右下方向に移動している．足立区をはじめ文京区，港区，江戸川区も右下方向に移動している．

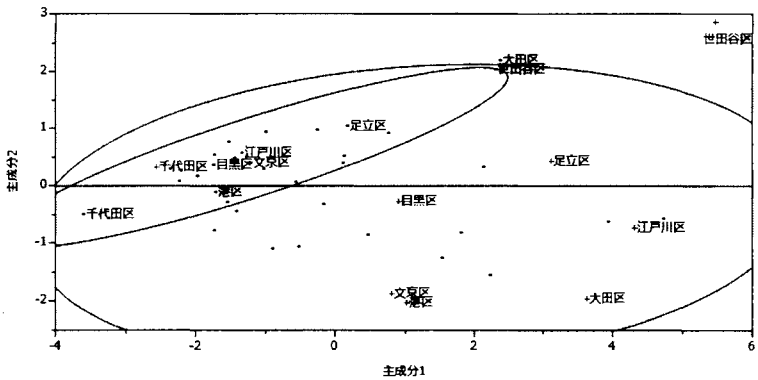
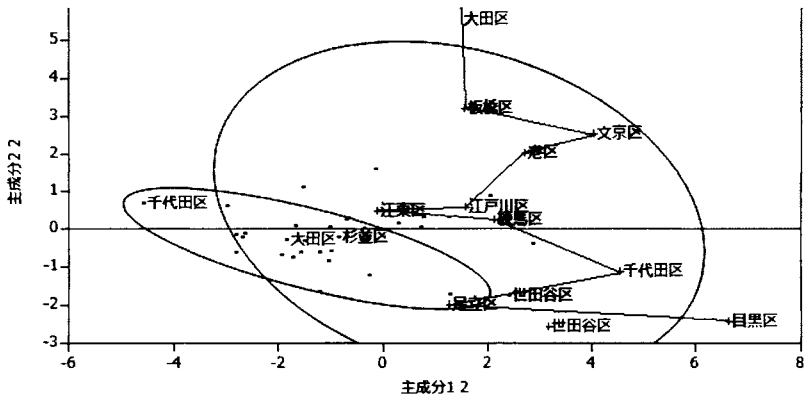


図 3　7 変数の主成分分析

図 4 は，5 入力 2 出力で用いた変数で 10 個の入出力の比（表 7 の最初の 10 変数）で主成分分析した結果である．そして 1986 年と 2011 年のデータを個別に 95%正規確率楕円を描いた．左下の小さなものが 1986 年で，右の大きな正規確率楕円が 2011 年である．また 46 図書館の CCR モデルで選ばれた手本の 10 図書館を線分で結んだ．凸包になっていないが効率的フロンティアに対応している．1986 年の 95%正規確率楕円が効率的フロンティアに含まれている．また 1986 年に 7 区ある手本のうちの杉並区は第 3 象限に，世田谷区は千代田区と足立区を結ぶ線分上にあり 1986 年から 2011 年にかけて効率的フロンティアは拡大したことを示す．

千代田区は，2 象限から 2011 年には第 4 象限の効率的フロンティアまで大きく動いた．図 3 と異なり変化率が大きかったことを示す．1986 年に効率的であった世田谷区は右下方向に業務を拡大しているが，1986 年の図書館業務の規模が大きかったので変化率は小さい．これに対して，規模の大きな大田区は右上に大きく変化しているのは職員数の未記入の問題が影響している．



16

図 4　主成分分析によるスコアプロットと 46 図書館の効率的フロンティア

5. まとめ

本研究では 1986 年と 2011 年の東京都 23 区の公立図書館の比較を，5 入力 2 出力モデルで行うことで，次のことが分かった．

・1986 年と 2011 年の 46 図書館を 5 入力 2 出力の CCR モデルで分析すると，手本は 2011 年単独の手本と同じ 10 図書館で，Inverted CCR モデルでは 1986 年単独の場合と同じく 3 図書館が非効率な手本になった．このことから，1986 年から 2011 年にかけて効率的フロンティアが拡大したことが分かる．

・1986 年で最も非効率であった千代田区は，25 年間で予算と人口は 1.04 倍と 1.03 倍しか増えていない．床面積，蔵書数，職員数は 1.66 倍，1.82 倍，2.02 倍増えたが，それ以上に貸出数と登録者数を 7.71 倍と 12.57 倍と図書館業務を大きく改善し，2011 年には手本になった．他の公立図書館も，図書館業務を大きく改善したものが多い．これは，図書館業務の外部委託化と図書館の複合センター化などで達成できたと考えられる．しかし，近年多くの組織で外注化が行われているが，以下の質的改善が同時に行われた結果といえる．

・区の直轄事業であれば司書を正式に採用できないが，専門業者であれば可能になる．

・公立図書館の 22 時までの開館や図書の 24 時間返却ポストの開設，託児サービスなどにより，勤労者や児童保護者の図書館利用層の積極的な拡大を行った．

・大学図書館や他自治体やグループへの積極的な貸し出しを行った．

・千代田区にみるように，神田の古書店などの地域との連携が行われ，単なる図書の貸し出しから従来の図書館の枠を超えた情報の提供という旧来の図書館業務には見られない試みもあり評価に値する．

また本研究事例は，DEA の評価における可視化と公平性を示す成功事例と考える．1986 年には DEA 効率値が 0.19 であった千代田区が 2011 年には目黒区と同じく手本になった点である．これまでも比率を用いた評価は単純でわかりやすいが個々で異なった結果になり，説得性に乏しかった．それを複数の入力と出力を重みで総合化し，評価対象に最適な DEA 効率値（$^{t}b_i*y_i/{}^{t}a_i*x_i$）を提案した点である．これによって規模の小さな千代田区や，予算や人口が少なく出力も 6 位と 7 位という目黒区が総合して効率的であることが分かった．さらに「1 入力固定改善法」で，複数の代替案を総合的に判断することで，評価対象の問題点が発見できた．

これまでの評価法であれば，千代田区や目黒区を評価する基準がなく，図書館業務の規模の大きな大田区や世田谷区に隠れてクローズアップされなかったであろう．日本では企業の事業部評価においても，採算性は良いが規模の小さい事業部を評価することは一般的に行われていない．それが DEA 効率値という尺度でもって初めて正しく評価できる道が開かれた点は大きい．今後企業の経営効率性の分析や，経営効率性になじまない大学経営にも利用できる可能性がある．

謝辞：本研究内容に関して，2011 年の調査を担当された東京都立中央図書館の担当者の方に説明会を開催し，職員数などに関して意見交換する機会を得たことに感謝する．

参考文献

[1] J.P.Sall, L.Creighton & A.Lehman(2004).JMP を用いた統計およびデータ分析入門（第 3 版）. SAS Institute Japan ㈱.［新村秀一監修］.

[2]L.Schrage(2003). Optimizer Modeling with LINGO. LINDO Systems Inc.

[3]S.Shinmura(2012).Relationship between the DEA cluster and the lower limit of weights. Proceedings of DEA Symposium 2012, pp.63-68.

[4] K.Tone(1988).Introduction to Efficiency Analysis of a company-DEA (1). Operations Research, 32/12.

[5]新村秀一(2004). JMP 活用　統計学とっておき勉強法. 講談社.

[6] 新村秀一(2007). Excel と LINGO で学ぶ数理計画法. 丸善.

[7] 新村秀一(2011).数理計画法による問題解決法. 日科技連出版社.

[8] _(2011).DEA による回帰型データのクラスター分析. 成蹊大学一般研究報告, 45/3, 1-37.

[9] 刀根薫(1993).経営効率性の測定と改善 - 包絡分析法 DEA による-. 日科技連出版社.

[10] _(1987).DEA 事例集. Institute for Policy Science Research Report.1-33.

[11] 東京都中央図書館.「2011 年東京都公立図書館調査」. 1-138.

LS-Means 再考

– GLM と PLM によるモデル推定後のプロセス –

○魚住 龍史
京都大学大学院医学研究科 医学統計生物情報学

The current innovation for LS-Means: implementation by using both GLM and PLM procedures

Ryuji Uozumi
Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine

要旨

SAS では，線形モデルによる解析を行うためのプロシジャが多く実装されている．これらのプロシジャでは，モデルの推定後にパラメータの線形式に対する推定が必要となることがある．例えば，LSMEANS ステートメントによる各群の LS-Means「最小 2 乗平均」を算出し，比較することが挙げられる．LSMEANS ステートメントや ESTIMATE ステートメント等の機能は，SAS/STAT V9.22 から大幅に拡張され，より多くのプロシジャで実行できるようになった．さらに，上述したモデル推定後のプロセスは，以前のバージョンではプロシジャの実行とともに記述する必要があったが，SAS/STAT V9.22 から，モデル情報をアイテムストアとして保存するための STORE ステートメント，及びアイテムストアを呼び出すための PLM プロシジャが追加された．本稿では，GLM プロシジャを用いた，LSMEANS ステートメントによる LS-Means の算出について，各パラメータに対する係数を ESTIMATE ステートメントで明示的に概説する．次に，オプション機能として，OBSMARGINS オプション，BYLEVEL オプション，AT オプションを用いた LSMEANS ステートメントによる LS-Means の算出について解説する．また，GLM プロシジャによるモデル情報から，PLM プロシジャを用いて，LSMESTIMATE ステートメントによる推定を行う実行手順を示す．

キーワード：LS-Means，GLM，ESTIMATE，OBSMARGINS，BYLEVEL，AT，モデル情報，PLM，LSMESTIMATE

1 はじめに

SAS では，GLM プロシジャや LOGISTIC プロシジャなどのように，線形モデルによる解析を行うためのプロシジャが多く実装されている．これらのプロシジャでは，MODEL ステートメントにおいて推定するモデル式を記述し，SOLUTION オプションによってパラメータ推定値を表示することができる．さらに，データ解析では，モデル推定後のプロセスとして，パラメータの線形式に対する推定が必要となることがある．

例えば，LSMEANS ステートメントによる各群の最小 2 乗平均 (least squares means，以下，LS-Means) を算出し，比較することが挙げられる．LSMEANS ステートメントは，OBSMARGINS (OM) オプション，BYLEVEL オプション，AT オプションといったオプション機能がサポートされており，大変簡便かつ有用である．また，CONTRAST ステートメントや ESTIMATE ステートメントによる，任意の係数を用いた推定も考えられる． LSMEANS ステートメントの機能は，以前のバージョンにおいては GLM プロシジャなどの一部のプロシジャのみで利用できるものであった．しかし， SAS/STAT V9.22 から大幅に拡張され，表 1 のように，LOGISTIC プロシジャや PHREG プロシジャなどのより多くのプロシジャで実行できるようになった．また，新たに LSMESTIMATE ステートメントが追加されており，LS-Means の線形式に対する推定，検定を行うことができる．LSMESTIMATE ステートメントは，LSMEANS ステートメントと異なり，各 LS-Mean の値は算出されない．さらに，モデルパラメータの線形式を用いる ESTIMATE ステートメントと異なり，LSMESTIMATE ステートメントでは LS-Means の線形式を用いるため，より解釈しやすいステートメントといえる．本ユーザー総会においては，浜田 (2013) によって，PHREG プロシジャによる上記のステートメントの解説が行われ，決まりきった対比較等の不必要な項目まで冗長に出力されてしまう LSMEANS ステートメントに比べて，特定の群間比較のみ出力できる LSMESTIMATE ステートメントが推奨された [7].

表 1：各プロシジャにおけるステートメントの使用可否 [3]

プロシジャ	LSMEANS	CONTRAST	ESTIMATE	LSMESTIMATE
GLM	○	○	○	–
MIXED	○	○	○	☆
GENMOD	☆	○	○	☆
GLIMMIX	○	○	○	☆
LOGISTIC	☆	○	☆	☆
PHREG	☆	○	☆	☆
ORTHOREG	☆	–	☆	☆
PLM	☆	–	☆	☆
SURVEYREG	☆	○	○	☆
SURVEYLOGISTIC	☆	○	☆	☆
SURVEYPHREG	☆	–	☆	☆

○: 以前のバージョンから利用可能，☆: 新たに追加あるいは更新，–: サポートされていない

その他にも，SAS/STAT V9.22 から拡張された機能として，モデル情報をアイテムストアとして保存するための STORE ステートメント，及びアイテムストアを呼び出すための PLM プロシジャが追加された．上述した LS-Means を算出するモデル推定後のプロセスは，以前のバージョンではプロシジャの実行とともに記述する必要があり，オブザベーション数が多い場合やモデル式が複雑である場合，モデル推定に再度多くの時間を要することがあった．しかし，PLM プロシジャを用いると，推定されたモデル情報を呼び出し，統計量の算出，ODS GRAPHICS によるグラフ表示等を実施でき，再度の入力データセットの参照，線形モデルプロシジャの実行が不要であり，大変有用であるといえる．

本稿では，GLM プロシジャを用いた，LSMEANS ステートメントによる LS-Means の算出について，各パラメータに対する係数を ESTIMATE ステートメントで明示的に概説する．次に，オプション機能として，OM オプション，BYLEVEL オプション，AT オプションを用いた LSMEANS ステートメントによる LS-Means の算出について述べる．さらに，GLM プロシジャによるモデル情報を用いて，PLM プロシジャを実行する手順を述べ，GLM プロシジャではサポートされていない LSMESTIMATE ステートメントによる推定を行う手順を示す．

2 GLM プロシジャによる LS-Means の算出

2.1 GLM プロシジャによる LS-Means の数理

GLM プロシジャによる LS-Means の算出までのプロセスを示すために，表 2 の 2 元配置アンバランスデータを考える．

表 2 : 2 元配置アンバランスデータ例 (データセット名: twoway)

		Block	
		ブロック 1	ブロック 2
Treatment	治療 1	17. 19. 19. 21. 22. 28 ($n_{11}=6$)	30. 39. 43 ($n_{12}=3$)
	治療 2	22. 26. 30. 31. 33. 46 ($n_{21}=6$)	26. 29. 31. 33 ($n_{22}=4$)

このとき，切片パラメータを μ，治療効果を表すパラメータを α_i $(i=1,2)$，ブロック効果を表すパラメータを β_j $(j=1,2)$，治療 i とブロック j の交互作用効果を表すパラメータを γ_{ij}，治療 i，ブロック j における k 番目の観測パラメータを Y_{ijk} $(k=1,2,...,n_{ij})$，誤差パラメータを ε_{ijk} (互いに独立に平均 0，分散 σ^2 の正規分布に従うと仮定) とすると，

$$Y_{ijk}=\mu+\alpha_i+\beta_j+\gamma_{ij}+\varepsilon_{ijk} \tag{1}$$

と表せ，治療 i，ブロック j における母平均パラメータは

$$E[Y_{ij}]=\mu+\alpha_i+\beta_j+\gamma_{ij}$$

と表せる．また，治療 i，ブロック j の周辺母平均パラメータはそれぞれ

$$E[Y_{i\cdot}]=\frac{E[Y_{i1}]+E[Y_{i2}]}{2}=\mu+\alpha_i+\frac{\beta_1+\beta_2}{2}+\frac{\gamma_{i1}+\gamma_{i2}}{2}=\mu+\alpha_i+\bar{\beta}_{\cdot}+\bar{\gamma}_{i\cdot} \tag{2}$$

$$E[Y_{\cdot j}]=\frac{E[Y_{1j}]+E[Y_{2j}]}{2}=\mu+\frac{\alpha_1+\alpha_2}{2}+\beta_j+\frac{\gamma_{1j}+\gamma_{2j}}{2}=\mu+\bar{\alpha}_{\cdot}+\beta_j+\bar{\gamma}_{\cdot j}$$

と表記できる．

さらに，治療効果の差の推定を考えると，式 (2) より，

$$E[Y_{1\cdot}]-E[Y_{2\cdot}]=\alpha_1-\alpha_2+\frac{\gamma_{11}+\gamma_{12}-\gamma_{21}-\gamma_{22}}{2} \tag{3}$$

となり，γ_{ij} が含まれることに留意しなければならない．

ここで，パラメータベクトル$\boldsymbol{\beta}$を

$$\boldsymbol{\beta} = (\mu \quad \alpha_1 \quad \alpha_2 \quad \beta_1 \quad \beta_2 \quad \gamma_{11} \quad \gamma_{12} \quad \gamma_{21} \quad \gamma_{22})^T,$$

係数行列$\mathbf{L}$を

$$\mathbf{L} = \begin{bmatrix} 1 & 1 & 0 & 0.5 & 0.5 & 0.5 & 0.5 & 0 & 0 \\ 1 & 0 & 1 & 0.5 & 0.5 & 0 & 0 & 0.5 & 0.5 \\ 1 & 0.5 & 0.5 & 1 & 0 & 0.5 & 0 & 0.5 & 0 \\ 1 & 0.5 & 0.5 & 0 & 1 & 0 & 0.5 & 0 & 0.5 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{4}$$

として，(周辺) 母平均パラメータベクトルは$\mathbf{L}\boldsymbol{\beta}$を計算することによって求まる．LS-Means は，推定パラメータベクトル$\hat{\boldsymbol{\beta}}$から求まる$\mathbf{L}\hat{\boldsymbol{\beta}}$である．

2.2 LSMEANS ステートメントによる実行

GLM プロシジャでは，LSMEANS ステートメントを用いて，LS-Means を容易に算出することができる．表 2 のデータに対して，モデル式 (1) を用いて，LS-Means を求めるための SAS プログラム例及び実行結果を一部抜粋したものを図 1 に示す．

LSMEANS ステートメントを用いたプログラム

```
proc glm data=twoway;
  class treatment block;  model y=treatment | block;
  lsmeans treatment / cl stderr e; lsmeans block / cl stderr e;
  lsmeans treatment*block /  l stderr e;
run;quit;
```

係数ベクトルの出力 (治療効果)

最小 2 乗平均 Treatment の係数 効果	Treatment 水準 1	2
Intercept	1	1
Treatment 1	1	0
Treatment 2	0	1
Block 1	0.5	0.5
Block 2	0.5	0.5
Treatment*Block 1 1	0.5	0
Treatment*Block 1 2	0.5	0
Treatment*Block 2 1	0	0.5
Treatment*Block 2 2	0	0.5

LS-Means 算出結果

Treatment	Block	y の最小 2 乗平均	標準誤差
1	.	29.167	2.091
2	.	30.542	1.909
.	1	26.167	1.707
.	2	33.542	2.259
1	1	21.000	2.415
1	2	37.333	3.415
2	1	31.333	2.415
2	2	29.750	2.957

図 1 : GLM プロシジャによる LS-Means の算出 (1)

LS-Means の計算は，CLASS ステートメント及び MODEL ステートメントによるモデル推定後のプロセスとなるため，GLM プロシジャ内の LSMEANS ステートメントは上記 2 つのステートメントの後に記述する．

図 1 の例では，交互作用効果を含んだモデル式 (1) から得られる治療効果，ブロック効果，交互作用効果の LS-Means をそれぞれ求めている．なお，LSMEANS ステートメントの E オプションは，推定に用いる係数ベクトルを出力させるための機能である．推定に用いる係数ベクトルの出力は，式 (2) に対応する場合のみ抜粋して示している ($i = 1, 2$)．すなわち，式 (4) の $\mathbf{L}$ における第 1 行，第 2 行が該当する．

なお，表 1 に示すように，SAS/STAT V9.22 以降，LSMEANS ステートメントは多くのプロシジャでサポートされている．

2.3 ESTIMATE ステートメントによる実行

図 1 のように，LSMEANS ステートメントの E オプションを指定すれば，式 (4) の $\mathbf{L}$ のような係数 (行列) を出力させることは可能である．しかし，SAS プログラム上で明示的に確認はできない．そこで，ESTIMATE ステートメントを用いて，各パラメータに対する係数を指定した上で，LS-Means を求めるための SAS プログラム例及び実行結果を一部抜粋したものを図 2 に示す．

ESTIMATE ステートメントを用いたプログラム

```
proc glm data=twoway;
  class treatment block;  model y=treatment | block;
  estimate "Treatment 1" intercept 1 treatment 1 0 block 0.5 0.5 treatment*block 0.5 0.5 0 0 / e;
  estimate "Treatment 2" intercept 1 treatment 0 1 block 0.5 0.5 treatment*block 0 0 0.5 0.5 / e;
  estimate "Block 1" intercept 1 treatment 0.5 0.5 block 1 0 treatment*block 0.5 0 0.5 0 / e;
  estimate "Block 2" intercept 1 treatment 0.5 0.5 block 0 1 treatment*block 0 0.5 0 0.5 / e;
  estimate "Treatment 1 Block 1" intercept 1 treatment 1 0 block 1 0 treatment*block 1 0 0 0 / e;
  estimate "Treatment 1 Block 2" intercept 1 treatment 1 0 block 0 1 treatment*block 0 1 0 0 / e;
  estimate "Treatment 2 Block 1" intercept 1 treatment 0 1 block 1 0 treatment*block 0 0 1 0 / e;
  estimate "Treatment 2 Block 2" intercept 1 treatment 0 1 block 0 1 treatment*block 0 0 0 1 / e;
run;quit;
```

係数ベクトルの出力 (治療効果)

推定 Treatment 1 の係数	行 1
Intercept	1
Treatment 1	1
Treatment 2	0
Block 1	0.5
Block 2	0.5
Treatment*Block 1 1	0.5
Treatment*Block 1 2	0.5
Treatment*Block 2 1	0
Treatment*Block 2 2	0

推定 Treatment 2 の係数	行 1
Intercept	1
Treatment 1	0
Treatment 2	1
Block 1	0.5
Block 2	0.5
Treatment*Block 1 1	0
Treatment*Block 1 2	0
Treatment*Block 2 1	0.5
Treatment*Block 2 2	0.5

LS-Means 算出結果

パラメータ	*推定値*	*標準誤差*	*t 値*	*Pr > |t|*
Treatment 1	29.167	2.091	13.95	<.0001
Treatment 2	30.542	1.909	16.00	<.0001
Block 1	26.167	1.707	15.33	<.0001
Block 2	33.542	2.259	14.85	<.0001
Treatment 1 Block 1	21.000	2.415	8.70	<.0001
Treatment 1 Block 2	37.333	3.415	10.93	<.0001
Treatment 2 Block 1	31.333	2.415	12.98	<.0001
Treatment 2 Block 2	29.750	2.957	10.06	<.0001

図 2 : GLM プロシジャによる LS-Means の算出 (2)

ESTIMATE ステートメントにおいても E オプションを指定でき，推定に用いる係数ベクトルを出力させることができる．図 1 と同様に，推定に用いる係数ベクトルの出力は，式 (2) に対応する場合のみ抜粋して示している ($i = 1, 2$). 図 2 における LS-Means の結果は，LSMEANS ステートメントを用いて求めた図 1 の結果と一致していることを確認できる．

なお，LSMEANS ステートメント同様、ESTIMATE ステートメントも多くのプロシジャでサポートされている．しかし，LOGISTIC プロシジャ等では，CLASS ステートメントの PARAM オプションを指定でき，プロシジャによってデフォルトが異なる．そのため，ESTIMATE ステートメントにおける係数の指定に注意が必要である. 図 2 のような LS-Means の推定を行う場合, GLM 法によるデザイン行列の指定がわかりやすい [8].

3 GLM プロシジャによるオプション機能を用いた LS-Means の算出

同様の表 2 のデータを用いて, LSMEANS ステートメントのオプション機能を考える. 本セクションでは, モデル式 (1) よりシンプルな交互作用効果を含まないモデル式 (5) を用いて考える．

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \tag{5}$$

このとき，治療 i，ブロック j に対して，LSMEANS ステートメントによって求まる LS-Means はそれぞれ

$$E[Y_{i\cdot}] = \mu + \alpha_i + \frac{\beta_1 + \beta_2}{2} = \mu + \alpha_i + \bar{\beta}_\cdot \tag{6}$$

$$E[Y_{\cdot j}] = \mu + \frac{\alpha_1 + \alpha_2}{2} + \beta_j = \mu + \bar{\alpha}_\cdot + \beta_j$$

と表記できる．

さらに，治療効果の差の推定を考えると，式 (6) より，

$$E[Y_{1\cdot}] - E[Y_{2\cdot}] = \alpha_1 - \alpha_2 \tag{7}$$

となり，式 (3) と異なり，α_i のみで表せる．

パラメータベクトル $\boldsymbol{\beta}$ を

$$\boldsymbol{\beta} = (\mu \quad \alpha_1 \quad \alpha_2 \quad \beta_1 \quad \beta_2)^T,$$

係数行列 $\mathbf{L}$ を

$$\mathbf{L} = \begin{bmatrix} 1 & 1 & 0 & 0.5 & 0.5 \\ 1 & 0 & 1 & 0.5 & 0.5 \\ 1 & 0.5 & 0.5 & 1 & 0 \\ 1 & 0.5 & 0.5 & 0 & 1 \end{bmatrix} \tag{8}$$

とすると，LS-Means は，推定パラメータベクトル $\hat{\boldsymbol{\beta}}$ から求まる $\mathbf{L}\hat{\boldsymbol{\beta}}$ である．

3.1 OM オプションの利用

治療効果に対する LS-Means を考えると，表 2 のデータのようにブロックが 2 水準の場合，β_1, β_2 に対する係数 c_1, c_2 は

$$c_1 = 1/2 = 0.5, \quad c_2 = 1/2 = 0.5 \qquad \left(\sum_k c_k = 1\right)$$

となり，各水準に対して等しい割合が係数として用いられていることがわかる．

一方，LSMEANS ステートメントの OBSMARGINS (OM) オプションを用いると，入力データセットにおける各水準のオブザベーション数に依存して，係数が設定される．表 2 のデータでは，各ブロックにおけるデータ数は

$$n_{\cdot 1} = \sum_i n_{i1} = 12$$

$$n_{\cdot 2} = \sum_i n_{i2} = 7$$

であるため，係数c_1, c_2は

$$c_1 = n_{\cdot 1} / n_{\cdot\cdot} = \sum_i n_{i1} / \sum_i \sum_j n_{ij} = 12/19 = 0.6316$$

$$c_2 = n_{\cdot 2} / n_{\cdot\cdot} = \sum_i n_{i2} / \sum_i \sum_j n_{ij} = 7/19 = 0.3684$$

となる．

すなわち，OM オプションを用いると，治療i の LS-Means は

$$E[Y_{i\cdot}] = \mu + \alpha_i + \frac{12}{19}\beta_1 + \frac{7}{19}\beta_2$$

と表記でき，治療効果の差の LS-Means は

$$E[Y_{1\cdot}] - E[Y_{2\cdot}] = \alpha_1 - \alpha_2 \tag{9}$$

となり，式 (7) と同様にα_iのみで表せる．係数行列を考えると，式 (8) の代わりに

$$\mathbf{L} = \begin{bmatrix} 1 & 1 & 0 & 0.6316 & 0.3684 \\ 1 & 0 & 1 & 0.6316 & 0.3684 \\ 1 & 0.4737 & 0.5263 & 1 & 0 \\ 1 & 0.4737 & 0.5263 & 0 & 1 \end{bmatrix}$$

を用いて，$\mathbf{L}\hat{\boldsymbol{\beta}}$を計算している．

LSMEANS ステートメントの OM オプションを利用して，モデル式 (5) から LS-Means を求めるための SAS プログラム例，及び対応する ESTIMATE ステートメントによる SAS プログラム例を図 3 に示す．

図 3 の SAS プログラムを実行すると，LSMEANS ステートメント及び ESTIMATE ステートメントにより求めた LS-Means は一致していることが確認できる．

さらに，治療効果の差の推定を考えると，式 (9) のように表せ，交互作用効果を含まないモデルでは，OM オプションの使用の有無に関わらず，同様の結果となる．OM オプションの有無で LS-Means の結果を比較するための SAS プログラム例及び実行結果を一部抜粋したものを図 4 に示す．

図 4 の SAS プログラムを実行すると，LSMEANS ステートメントにより求めた LS-Means は，治療効果の差の推定に関しては一致していることが確認できる．しかし，各治療の LS-Means は異なった結果となる．

LSMEANS ステートメント (OM オプション) を用いたプログラム

```
proc glm data=twoway;
  class treatment block; model y=treatment block;
  lsmeans treatment / om cl stderr e;
run;quit;
```

ESTIMATE ステートメントを用いたプログラム

```
proc glm data=twoway;
  class treatment block; model y=treatment block;
  estimate "Treatment 1" intercept 1 treatment 1 0 block 0.6316 0.3684 / e;
  estimate "Treatment 2" intercept 1 treatment 0 1 block 0.6316 0.3684 / e;
run;quit;
```

図 3 : GLM プロシジャによる LS-Means の算出 (3)

OM オプションなし (デフォルト)

```
proc glm data=twoway;
  class treatment block; model y=treatment block;
  lsmeans treatment / cl diff e;
run;quit;
```

Treatment y の最小 2 乗平均　95% 信頼限界

1	27.538	22.166　32.910
2	31.356	26.343　36.369

効果 Treatment 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	-3.818	-11.038	3.402

OM オプションあり

```
proc glm data=twoway;
  class treatment block; model y=treatment block;
  lsmeans treatment / om cl diff e;
run;quit;
```

Treatment y の最小 2 乗平均　95% 信頼限界

1	26.675	21.443　31.907
2	30.493	25.530　35.456

効果 Treatment 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	-3.818	-11.038	3.402

図 4 : GLM プロシジャによる LS-Means の算出 (4)

3.2 OM BYLEVEL オプションの利用

表 2 のデータに対して，LSMEANS ステートメントの OM オプションに加えて，BYLEVEL オプションも用いると，入力データセットにおける各水準のオブザベーション数に依存して，以下のように係数が設定される．

【治療 1】　$c_1 = n_{11}/n_{1\cdot} = n_{11}/\sum_j n_{1j} = 6/9 = 0.6667$

$c_2 = n_{12}/n_{1\cdot} = n_{12}/\sum_j n_{1j} = 3/9 = 0.3333$

【治療 2】　$c_1 = n_{21}/n_{2\cdot} = n_{21}/\sum_j n_{2j} = 6/10 = 0.6$

$c_2 = n_{22}/n_{2\cdot} = n_{22}/\sum_j n_{2j} = 4/10 = 0.4$

OM オプションのみを使用した場合，各治療に対して共通の係数c_1, c_2を用いた．BYLEVEL オプションも加えると，各治療に対する係数をそれぞれ計算する．

すなわち，OM オプションを用いると，治療i の LS-Means はそれぞれ

$$E[Y_{1\cdot}] = \mu + \alpha_1 + \frac{6}{9}\beta_1 + \frac{3}{9}\beta_2$$

$$E[Y_{2\cdot}] = \mu + \alpha_2 + \frac{6}{10}\beta_1 + \frac{4}{10}\beta_2$$

と表記でき，治療効果の差の LS-Means は

$$E[Y_{1\cdot}] - E[Y_{2\cdot}] = \alpha_1 - \alpha_2 + \left(\frac{6}{9} - \frac{6}{10}\right)\beta_1 + \left(\frac{3}{9} - \frac{4}{10}\right)\beta_2 \tag{10}$$

となり，式 (7) や式 (9) と異なり，α_iのみで表せない．係数行列を考えると，式 (8) の代わりに

$$\mathbf{L} = \begin{bmatrix} 1 & 1 & 0 & 0.6667 & 0.3333 \\ 1 & 0 & 1 & 0.6 & 0.4 \\ 1 & 0.5 & 0.5 & 1 & 0 \\ 1 & 0.4286 & 0.5714 & 0 & 1 \end{bmatrix}$$

を用いて，$\mathbf{L}\hat{\boldsymbol{\beta}}$を計算している．

LSMEANS ステートメントの OM オプション及び BYLEVEL オプションを利用して，LS-Means を求めるための SAS プログラム例，及び対応する ESTIMATE ステートメントによる SAS プログラム例を図 5 に示す．

LSMEANS ステートメント (OM BYLEVEL オプション) を用いたプログラム

```
proc glm data=twoway;
  class treatment block; model y=treatment block;
  lsmeans treatment / om bylevel cl stderr e;
run;quit;
```

ESTIMATE ステートメントを用いたプログラム

```
proc glm data=twoway;
  class treatment block; model y=treatment block;
  estimate "Treatment 1" intercept 1 treatment 1 0 block 0.6667 0.3333 / e;
  estimate "Treatment 2" intercept 1 treatment 0 1 block 0.6 0.4 / e;
run;quit;
```

図 5 : GLM プロシジャによる LS-Means の算出 (5)

図 5 の SAS プログラムを実行すると，LSMEANS ステートメント及び ESTIMATE ステートメントにより求めた LS-Means は一致していることが確認できる．

さらに，OM オプション及び BYLEVEL オプションとデフォルトの LS-Means の結果を比較するための SAS プログラム例及び実行結果を一部抜粋したものを図 6 に示す．

OM オプションなし (デフォルト)

```
proc glm data=twoway;
  class treatment block; model y=treatment block;
  lsmeans treatment / cl diff e;
run;quit;
```

Treatment	y の最小 2 乗平均	95% 信頼限界	
1	27.538	22.166	32.910
2	31.356	26.343	36.369

効果 Treatment 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	-3.818	-11.038	3.402

OM BYLEVEL オプションあり

```
proc glm data=twoway;
  class treatment block; model y=treatment block;
  lsmeans treatment / om bylevel cl diff e;
run;quit;
```

Treatment	y の最小 2 乗平均	95% 信頼限界	
1	26.444	21.219	31.670
2	30.700	25.743	35.657

効果 Treatment 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	-4.256	-11.458	2.947

図 6 : GLM プロシジャによる LS-Means の算出 (6)

図 6 の SAS プログラムを実行すると，LSMEANS ステートメントにより求めた LS-Means は，OM オプション及び BYLEVEL オプションを利用すると，デフォルトと結果が一致しないことがわかる．治療効果の差の推定を考える場合，式 (10) のように表せ，式 (7) とは大きく異なる表記となるため，結果は一致しないことになる．

3.3 AT オプションの利用

モデル式 (4) に連続量 z_{ijk} を説明変数に含んだモデル (11) を考える．

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta \cdot z_{ijk} + \varepsilon_{ijk} \tag{11}$$

このとき，治療 i，ブロック j に対して，LSMEANS ステートメントによって求まる LS-Means は

$$E[Y_{i\cdot}] = \mu + \alpha_i + \frac{\beta_1 + \beta_2}{2} + \bar{z} \cdot \delta \tag{12}$$

$$E[Y_{j}] = \mu + \frac{\alpha_1 + \alpha_2}{2} + \beta_j + \bar{z} \cdot \delta$$

と表記でき，パラメータベクトル $\boldsymbol{\beta}$ を

$$\boldsymbol{\beta} = (\mu \quad \alpha_1 \quad \alpha_2 \quad \beta_1 \quad \beta_2 \quad \delta)^T,$$

係数行列 $\mathbf{L}$ を

$$\mathbf{L} = \begin{bmatrix} 1 & 1 & 0 & 0.5 & 0.5 & \bar{z} \\ 1 & 0 & 1 & 0.5 & 0.5 & \bar{z} \\ 1 & 0.5 & 0.5 & 1 & 0 & \bar{z} \\ 1 & 0.5 & 0.5 & 0 & 1 & \bar{z} \end{bmatrix}$$

とすると，LS-Means は，推定パラメータベクトル $\hat{\boldsymbol{\beta}}$ から求まる $\mathbf{L}\hat{\boldsymbol{\beta}}$ である．式 (12) におけるパラメータ δ の係数として，算術平均 $\bar{z}$ が用いられていることがわかる．

ここで，表2のデータは$\bar{z}=14.5263$であると仮定する．このとき，LSMEANSステートメントのATオプションを利用して，LS-Meansを求めるためのSASプログラム例，及び対応するESTIMATEステートメントによるSASプログラム例を図7に示す．

LSMEANSステートメント

```
proc glm data=twoway;
  class treatment block;
  model y=treatment block z;
  lsmeans treatment / cl stderr e;
run;quit;
```

ATオプション

```
proc glm data=twoway;
  class treatment block;
  model y=treatment block z;
  lsmeans treatment / cl stderr e at z=14.5263;
run;quit;
```

ESTIMATEステートメント

```
proc glm data=twoway;
  class treatment block;
  model y=treatment block z;
  estimate "Treatment 1"
           intercept 1 treatment 1 0
           block 0.5 0.5 z 14.5263 / e;
  estimate "Treatment 2"
           intercept 1 treatment 0 1
           block 0.5 0.5 z 14.5263 / e;
run;
quit;
```

図7：GLMプロシジャによるLS-Meansの算出 (7)

LSMEANSステートメントのATオプションを用いると，図7における "at z=14.5263" の値を変えることによって，$\bar{z}$の代わりに任意の値を係数として用いることができる．

4 PLMプロシジャによるLS-Meansの算出

前節まで，モデル推定後のLS-Meansの算出までのプロセスについて，GLMプロシジャ内の記述方法を示した．OMオプションのようなオプションを用いた結果が必要となった場合，オプションを追記した上で，再びGLMプロシジャを実行しなければならない．このような操作は，オブザベーション数が多い場合やモデル式が複雑である場合，モデル推定に再度多くの時間を要することになる．

4.1 PLMプロシジャの概要

SAS/STAT V9.22より，GLMプロシジャ等のプロシジャを用いて推定されているモデル情報を呼び出し，統計量やグラフ表示などを行うことのできる，PLMプロシジャが追加された．

PLMプロシジャを利用するにあたっては，まずGLMプロシジャ等のプロシジャを実行する際にSTOREステートメントを用いて，プロシジャから得られるモデル情報，すなわちアイテムストアをバイナリファイルとして保存する．STOREステートメントは，表1に示したプロシジャで実行できる．そして，PLMプロシジャのRESTOREステートメントを用いて，アイテムストアとして保存したモデル情報を呼び出し，統計量やグラフ表示などを行うことができる．モデル情報に基づく実行であるため，再度の入力データセットの参照，線形モデルのプロシジャを実行することなく，モデル推定後のプロセスを実行できる．

4.2 LSMEANS ステートメントによる実行

GLM プロシジャによるモデル情報を "twowayfit" という名前のアイテムストアに保存し，PLM プロシジャで実行するまでの SAS プログラム例を図 8 に示す．表 2 のデータに対して，交互作用効果を含むモデル式 (1) による解析を考えている．

STORE ステートメントによるアイテムストアの保存

```
proc glm data=twoway;
  class treatment block; model y=treatment | block;
  store twowayfit;
run;
quit
```

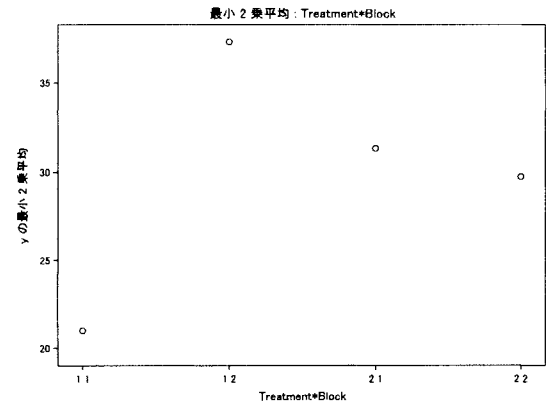
PLM プロシジャによる LS-Means の算出

```
proc plm restore=twowayfit;
  lsmeans treatment / cl;
  lsmeans block / cl;
  lsmeans treatment*block / cl;
run;
```

図 8 : GLM 及び PLM プロシジャによる LS-Means の算出 (1)

図 8 の SAS プログラムの実行結果のうち，治療とブロックの交互作用効果に対する LS-Means として，ODS GRAPHICS による出力結果を図 9 に示す．

GLM プロシジャによる出力



PLM プロシジャによる出力

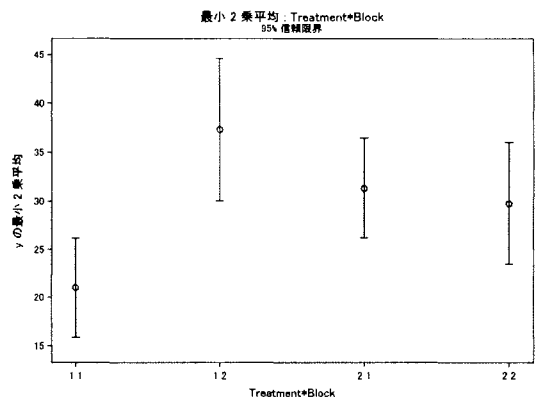


図 9 : GLM 及び PLM プロシジャによる LS-Means の算出 (2)

図 9 のように，GLM プロシジャでは各水準の LS-Means がプロットされる一方、PLM プロシジャでは信頼区間も出力される点で有用といえる．

なお，アイテムストアとして保存したモデル情報は，図 10 のような SAS プログラムを実行することによって確認できる．

```
proc plm restore=twowayfit;
   show all;
run;
```

図 10 : すべてのモデル情報の表示を行う SAS プログラム

4.3 LSMESTIMTATE ステートメントによる実行

SAS/STAT V9.22 より，LSMESTIMATE ステートメントが追加された．モデルパラメータの線形式を用いる ESTIMATE ステートメントと異なり，LSMESTIMATE ステートメントでは LS-Means の線形式を用いる．浜田 (2013) は，決まりきった対比較等の不必要な項目まで冗長に出力されてしまう LSMEANS ステートメントに比べて，特定の群間比較の LS-Means のみを出力できる LSMESTIMATE ステートメントを推奨している．しかし，表 1 で示したように，GLM プロシジャでは LSMESTIMATE ステートメントはサポートされていない．そこで，GLM プロシジャによるモデル情報を用いて，PLM プロシジャで LSMESTIMATE ステートメントによる推定を行う方法を示す．

図 8 において保存したアイテムストア "twowayfit" を用いて，PLM プロシジャによって LSMESTIMATE ステートメントによる SAS プログラム例を図 11 に示す．

図 11 のように，LSMESTIMATE ステートメントでは，LS-Means を算出したい変数とその係数のみを指定すれば，ESTIMATE ステートメントを用いて求めた図 2 と同様の結果を得ることができる．

LSMESTIMATE ステートメントを用いたプログラム

```
proc plm restore=twowayfit;
  lsmestimate treatment "Treatment 1" 1 0 / cl e;
  lsmestimate treatment "Treatment 2" 0 1 / cl e;
  lsmestimate block "Block 1" 1 0 / cl e;  lsmestimate block "Block 2" 0 1 / cl e;
  lsmestimate treatment * block "Treatment 1 Block 1" 1 0 0 0 / cl e;
  lsmestimate treatment * block "Treatment 1 Block 2" 0 1 0 0 / cl e;
  lsmestimate treatment * block "Treatment 2 Block 1" 0 0 1 0 / cl e;
  lsmestimate treatment * block "Treatment 2 Block 2" 0 0 0 1 / cl e;
run;
```

図 11 : GLM 及び PLM プロシジャによる LS-Means の算出 (3)

4.4 SLICE ステートメントによる実行

SLICE ステートメントは，LSMESTIMATE ステートメントと同様に，SAS/STAT V9.22 から追加された機能であり，カテゴリ変数が 2 次以上の交互作用項に対し，1 つの変数の水準をスライスした解析ができる．LSMEANS ステートメントと同じオプションを用いることができる．LSMESTIMATE ステートメントと同様に．GLM プロシジャではサポートされていない．

図 8 において保存したアイテムストア "twowayfit" を用いて，PLM プロシジャによって SLICE ステートメントによる SAS プログラム例及び出力結果を図 12 に示す．

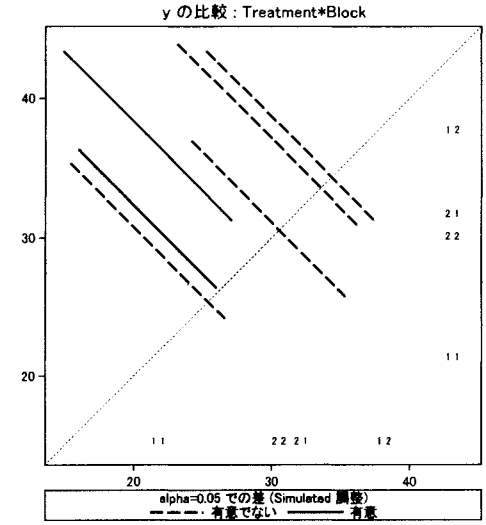
図 12 は，GLM プロシジャによる出力結果を比較対照として出力している．GLM プロシジャによる対比較の結果においては，ODS GRAPHICS が冗長に出力されてしまう一方，PLM プロシジャの SLICE ステートメントを用いると，特定の水準に対する交互作用効果の結果のみ出力できる．

GLM プロシジャ　LSMEANS ステートメント

```
proc glm data=twoway;
  class treatment block; model y=treatment | block;
  lsmeans treatment * block
    / diff cl adjust=simulate(report seed=4989);
run;quit;
```

PLM プロシジャ　SLICE ステートメント

```
proc plm restore=twowayfit;
  slice treatment * block
    / diff cl sliceby(block="1");
run;
```



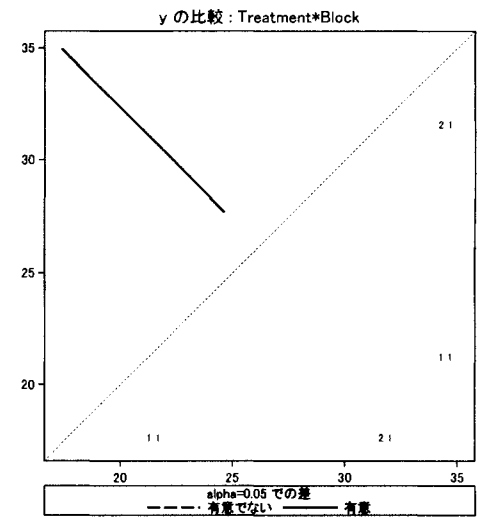


図 12 : GLM 及び PLM プロシジャによる LS-Means の算出 (4)

5 まとめ

本稿では,線形モデルによる解析を行うためのプロシジャによるモデルの推定後に算出する LS-Means について，LSMEANS ステートメントで算出される数理を示した上で，ESTIMATE ステートメントを用いて SAS プログラム上で明示的に概説した．次に，LSMEANS ステートメントのオプション機能として，OM オプション，BYLEVEL オプション，AT オプションを用いた LS-Means の算出方法を詳述した．さらに，線形モデルによる解析を行うためのプロシジャによって，推定されたモデル情報をアイテムストアとして保存した上で，PLM プロシジャによる推定を行う方法を解説した．加えて，LSMESTIMATE ステートメント及び SLICE ステートメントによる解析方法についても示した．

表 1 に示したように，現在の SAS のバージョンでは，LS-Means 算出のためのステートメントの選択肢は幅広く，多くのプロシジャで実行できる．特に，LSMEANS ステートメントは，ESTIMATE ステートメントよりもはるかに簡潔な SAS プログラムで LS-Means を算出できるといえる．さらに，OM オプションを指定すれば，入力データセットにおける各水準のオブザベーション数に依存して係数を設定できる．例えば，ランダム化臨床試験のデータ解析を考えた場合，割付因子として調整した共変量を説明変数に含めて，群間の LS-Means の差の推定を行うことが多い．このとき，LS-Means の差であれば，OM オプションを指定した結

果はデフォルトの結果と一致する．しかし，LS-Means の差ではなく，各群の LS-Means を求めることに関心がある場合，割付因子として調整している説明変数に対する係数を指定できる，LSMEANS ステートメントの OM オプションは有用であるといえる．また，非線形なデータに対して，ある特定の説明変数の値における LS-Means の推定を行う場合，LSMEANS ステートメントの AT オプションが有用になると考えられる．

参考文献

[1] Cai W. Making Comparisons Fair: How LS-Means Unify the Analysis of Linear Models. *Proceedings of the SAS Global Forum*. Cary, NC: SAS Institute Inc., 2014. Available at http://support.sas.com/resources/papers/proceedings14/SAS060-2014.pdf.

[2] High R. Plotting Differences among LSMEANS in Generalized Linear Models. *Proceedings of the SAS Global Forum*. Cary, NC: SAS Institute Inc., 2014. Available at http://support.sas.com/resources/papers/proceedings14/1902-2014.pdf.

[3] Kiernan K, Tobias R, Gibbs P, Tao J. Making CONTRAST and ESTIMATE Statements Made Easy: The LSMESTIMATE Statement. *Proceedings of the SAS Global Forum*. Cary, NC: SAS Institute Inc., 2011. Available at http://support.sas.com/resources/papers/proceedings11/351-2011.pdf.

[4] SAS Institute Inc. *SAS/STAT(R) 9.3 User's Guide*, Cary, NC, USA: SAS Institute Inc; 2011.

[5] 関根暁史. PLM プロシジャによる回帰分析と予測の分離. SAS ユーザー総会 論文集 2013, 275–290.

[6] 竹内啓, 高橋行雄, 大橋靖雄, 芳賀敏郎. SAS による実験データの解析. 東京大学出版会, 1989.

[7] 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.

[8] 吉田早織, 魚住龍史. 線形モデルにおける CLASS ステートメントの機能. SAS ユーザー総会 論文集 2014.

連絡先

E-mail : *uozumi@kuhp.kyoto-u.ac.jp*

SASによる二項比率の差の非劣性検定の比較

武藤彬正　宮島育哉　榊原伊織
株式会社タクミインフォメーションテクノロジー

Exact method of non-inferiority test for two binomial proportions using SAS

Akimasa Muto　Ikuya Miyajima　Iori Sakakibara
Takumi Information Technology Inc.

要旨

非劣性検定とは、新規の薬剤の治療効果が既存の薬剤の治療効果と比較して、ある程度以上は劣っていないことを示す検定である。特に二項比率の差の非劣性検定では、新規の薬剤の治療効果と既存の薬剤の治療効果の比率の差が、一定の評価基準(非劣性マージン)を下回らないことを示す事である。

SAS 9.4 では Wald, Hauck and Anderson, Farrington and Manning といった検定統計量が利用可能だが、漸近正規性により *p-value* を算出するため、例数が少ない場合には Type I Error を起こす確率が名目の有意水準を上回ることがある。一方で、例数が少ない場合でも名目の水準を上回らない二項分布の確率関数を直接的に用いた Exact な方法や、Exact な検定手法の計算効率を向上させた Exact Like な方法も提案されている。本研究では、これらの検定手法について紹介し比較するとともに、そのプログラムを提供する。

キーワード：非劣性検定　二項比率の差　Exact な方法　Exact Like な方法　Type I Error
FREQ プロシジャ　Noninf　RiskDiff

1. Introduction

非劣性検定とは、新規の薬剤の治療効果が既存の薬剤の治療効果に比べて、ある程度以上には劣っていないことを示す検定である。すなわち、新規の薬剤の治療効果 π_1 と既存の薬剤の治療効果 π_2 の差 $\delta = \pi_1 - \pi_2$ が、一定の評価基準非劣性マージン: $\Delta_0 (\Delta_0 > 0)$ を下回らないことを示す事である。

非劣性と関連する概念に、優越性という概念がある。優越性検定とは、新規の薬剤の治療効果が、既存の薬剤の治療効果に比べ優れていることを示す検定である。仮説検定と信頼区間の関連性を用いると、非劣性、優越性及び劣性の関係は次に示す Figure 1 のようになる。

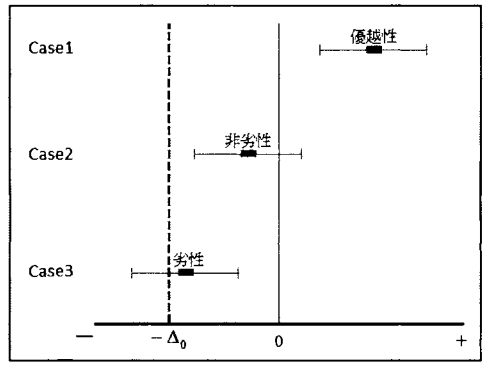


Figure 1　優越性と非劣性/劣性の関係

Figure1 における Case1 は、信頼下限が$-\Delta_0$と 0 の線を超えているので、新薬は既存薬に対し「優越性あり・非劣性あり」となる場合を示している。Case2 は、信頼下限が$-\Delta_0$の線を超えているが、0 の線を超えていないので「優越性なし・非劣性あり」となる場合を示している。また、Case3 は$-\Delta_0$と 0 の線のいずれも超えていないので「優越性なし・非劣性なし」すなわち劣性の場合を示している。

2. Notation and Method

X_1およびX_2は、それぞれ独立した二項分布に従う確率変数とする。X_1のサンプルサイズをn_1、二項比率をπ_1とし、$X_1 \sim B(n_1, \pi_1)$と示し、X_2のサンプルサイズをn_2、二項比率をπ_2とし、$X_2 \sim B(n_2, \pi_2)$と示す。非劣性検定での、帰無仮説と対立仮説は、非劣性マージンΔ_0を用いて次のように表わせる。

$$H_0 : \pi_1 - \pi_2 \leq -\Delta_0$$
$$H_1 : \pi_1 - \pi_2 > -\Delta_0$$

母比率の差を$\delta = \pi_1 - \pi_2$とし、標本比率の差を、$\hat{\delta} = \hat{\pi}_1 - \hat{\pi}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$とする。すなわち、帰無仮説での期待値は次の式で与えられる。

$$E(\hat{\delta}) = \pi_1 - \pi_2 = -\Delta_0$$

また、$\hat{\delta}$の分散は、

$$V(\hat{\delta}) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

である。

2.1. 近似的な手法

非劣性を統計的に確かめるには、非劣性マージンを含めた仮説検定、すなわち非劣性検定を行う必要がある。二項比率の差の検定を行う場合、新薬の標本比率 $\hat{\pi}_1$ と既存薬の標本比率 $\hat{\pi}_2$ の差に非劣性マージン Δ_0 を加え、それを標準偏差で除した

$$Z = \frac{(\hat{\pi}_1 - \hat{\pi}_2) + \Delta_0}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}}$$

が検定統計量となる。

ただし、標準偏差には未知のパラメーターである π_1 と π_2 が含まれている。未知のパラメーターの推定方法が複数あり、それに応じた検定統計量が提案されている。例えば、次のような検定統計量が提案されている。

2.1.1. Wald 検定統計量

π_1 と π_2 はそれぞれ独立しているとみなした場合、$\hat{\pi}_1$ と $\hat{\pi}_2$ がそれぞれ最尤推定量となる。すなわち、対立仮説のもとでの Wald 検定統計量は、次のように表せる。

$$Z_W = \frac{(\hat{\pi}_1 - \hat{\pi}_2) + \Delta_0}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}}$$

2.1.2. Farrington and Manning 検定統計量

Farrington and Manning(1990)で示された Farrington and Manning 検定統計量は未知のパラメーターである π_2 の推定に、帰無仮説のもとでの最尤推定量、即ち制約付き最尤推定値 $\tilde{\pi}_2$ 利用する統計量である。以下、Farrington and Manning 検定統計量を Z_F と表記する。帰無仮説の制約のもとで π_2 の最尤推定量は、次の尤度関数を用いて求めることができる。

$$P_{H_0}(X_1 = x_1, X_2 = x_2) = \binom{n_1}{x_1}\binom{n_2}{x_2}(\pi_2 + \Delta_0)^{x_1}(1 - \pi_2 - \Delta_0)^{n_1 - x_1}\pi_2^{x_2}(1-\pi_2)^{n_2 - x_2}$$

Z_F は、求めた $\tilde{\pi}_2$ を用いることで次のように表せる。

$$Z_F = \frac{(\hat{\pi}_1 - \hat{\pi}_2) + \Delta_0}{\sqrt{\frac{(\tilde{\pi}_2 - \Delta_0)(1 - \tilde{\pi}_2 + \Delta_0)}{n_1} + \frac{\tilde{\pi}_2(1 - \tilde{\pi}_2)}{n_2}}}$$

2.1.3. Hauck and Anderson 検定統計量

Hauck and Anderson (1986)で示された Hauck and Anderson 検定統計量は、上で述べてきた統計量と異なり、標準偏差に含まれる未知のパラメーターの推定を行うのではなく、連続調整を行い、近似効率を向上させたものである。以下、Hauck and Anderson 検定統計量を Z_H と表記する。

$$Z_H = \frac{\hat{\pi}_1 - \hat{\pi}_2 + \Delta_0 \pm cc}{\sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1 - 1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2 - 1}}}$$

ただし、$cc = \frac{1}{(2 \ \min(n_1, n_2))}$ である。

2.2. SAS に搭載されている検定手法

SAS9.4 に搭載されている検定手法は、Z_W、Z_F、Z_H を利用した検定手法である。これらの検定は FREQ プロシジャ内の TABLES ステートメントに RISKDIFF オプションを指定し、さらに RISKDIFF オプションに NONINF オプションを指定することで実行できる。

```
PROC FREQ DATA = 〈入力データ〉;
  tables 〈独立変数〉 * 〈従属変数〉 / riskdiff(noninf);
RUN;
```

デフォルトでは、検定統計量は、Z_W を利用し、非劣性マージンは 0.2 となっている。例えば、検定統計量を Z_F、非劣性マージンを 0.1 と設定する場合は次のような指定を行う。

```
PROC FREQ DATA = 〈入力データ〉;
 tables 〈独立変数〉 * 〈従属変数〉 / riskdiff(noninf
                                    margin = 0.1
                                    method = fm  );
RUN;
```

検定統計量は、METHOD=オプションで指定を行うことで、変更できる。Z_W ならば、[METHOD = WALD]、Z_F ならば[METHOD = FM]、Z_H ならば[METHOD = HA]と指定する。

2.3. Exact な方法

Exact な方法では、検定統計量を基準値として用い(本研究ではZ_Fを用いる)、帰無仮説のもとで取り得るすべての母比率について基準値より稀な観測値(x_1, x_2)の組み合わせをとる確率を算出する。得られた組み合わせごとの確率の内、上極限を$p-value$とする。$p-value$と有意水準を比較し、$p-value \leq \alpha$であれば帰無仮説は棄却される。

次に、帰無仮説より$\tilde{\pi}_1 - \tilde{\pi}_2 = -\Delta_0$とし、$\pi = \tilde{\pi}_2$とする。帰無仮説のもと$p-value$を算出する式を示す。

$$p-value^E = \sup_{\pi \in [\Delta_0, 1]} P_{H_0}\left(Z_F \leq Z_F(x_1, x_2) \mid \pi\right)$$
$$= \sup_{\pi \in [\Delta_0, 1]} \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} \binom{n_1}{x_1} \binom{n_2}{x_2} (\pi - \Delta_0)^{x_1} (1 - \pi + \Delta_0)^{n_1 - x_1} \pi^{x_2} (1-\pi)^{n_2 - x_2} I_{[Z_F \leq Z_F(x_1, x_2)]}$$

ただし、$I_{[Z_F \leq Z_F(x_1, x_2)]}$：定義関数とする。

2.4. Exact Like な方法

Kang and Chen(2000)で示された Exact Like な方法は、取り得るすべての母比率について計算する Exact な方法とは異なり、帰無仮説のもとで最尤推定値を用い、$p-value$を算出し検定を行う。帰無仮説のもとで、観測値(x_1^0, x_2^0)より推定されたπ_2の最尤推定量を$\tilde{\pi}$とし、$p-value$を算出する式を示す。

$$p-value^L = P_{H_0}\left(Z_F \leq Z_F(x_1^0, x_2^0) \mid \tilde{\pi}\right)$$
$$= \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} \binom{n_1}{x_1} \binom{n_2}{x_2} (\tilde{\pi} - \Delta_0)^{x_1} (1 - \tilde{\pi} + \Delta_0)^{n_1 - x_1} \tilde{\pi}^{x_2} (1-\tilde{\pi})^{n_2 - x_2} I_{[Z_F(x_1, x_2) \leq Z_F(x_1^0, x_2^0)]}$$

3. 検定方法の比較

SAS では、漸近正規性を用いた近似的な方法で、$p-value$を算出している。しかし近似的な方法以外にも、二項分布の確率関数を用いた Exact な方法、Exact な方法と近似的な方法を組み合わせた方法(以下、Exact Like な方法と呼ぶ) などの$p-value$を算出する方法が提案されている。

本研究において、検定手法の評価する指標は、Type I Error を起こす確率と、検出力とする。Type I Error を起こす確率は、安全性の指標であり、名目の有意水準αを超えずに近いことが望ましい。検出力は、感度の指標であり、高いほど望ましい。つまり、Type I Error を起こす確率がαを超えず、よりαに近いことに加え、検出力が高いことが、優れた統計手法の条件といえる。ただし、Type I Error を起こす確率が高ければ検出力が高くなり、Type I Error を起こす確率が低ければ検出力も低くなる傾向がある。

以下に、Z_W、Z_Fを用いた近似的な方法と検定統計量にZ_Fを用いたExactな方法、Exact Likeな方法について、Type I Errorを起こす確率、検出力を10000回のシミュレーションにより計算した結果を示す。Figure 2.1は$\Delta_0 = 0.1$とし、Figure 2.2は、$\Delta_0 = 0.25$とした結果である。

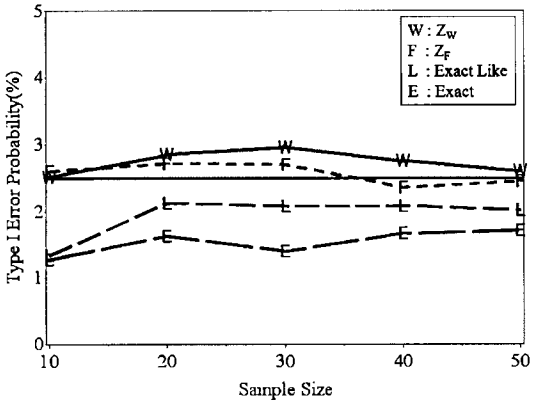


Figure 2.1. Type I Error を起こす確率
$n_1 = n_2$ and $\pi_1 = 0.7$ vs. $\pi_2 = 0.8$

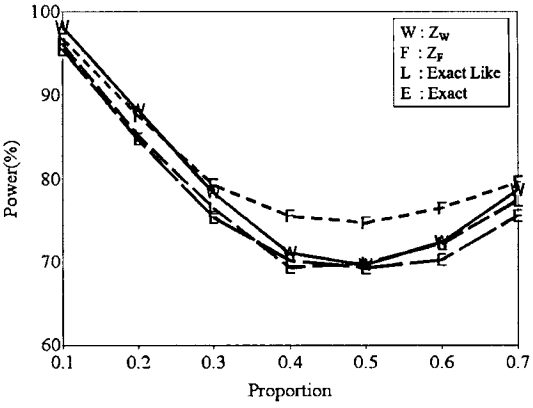


Figure 2.2. 検出力
$n_1 = n_2 = 50$ and $\pi_1 = \pi_2$

Figure 2.1は縦軸をType I Errorを起こす確率、横軸をnとするグラフである。また、αは2.5%とする。Z_Wを用いたWald検定などの近似法は、Type I Errorを起こす確率が度々α超える。しかし、Exactな方法では、αを超えることはない。Exact Likeな方法では、Type I Errorを起こす確率がαを超えることが少なく、Exactな方法より全体的にαに近い。

Figure 2.2.は縦軸を検出力、横軸を二項比率($\pi_1 = \pi_2$)とするグラフである。近似的な方法は、検出力が高い。Exact Likeな方法とExactな方法では、Exact Likeな方法の方が検出力高いことがわかる。

Type I Errorを起こす確率がαを超えることが少ないExact Likeな方法、αを超えないExactな方法はSASで提供されていない。そこで本研究は、この2つの検定手法について詳細を述べた後、そのプログラムを提供する。

4. プログラム

4.1. － Exactな方法、及びExact Likeな方法 －

```
%MACRO Noninf_EXL(N      =  /* 新薬のn数   */, x      = /* 新薬の有効数    */
                 ,M      =  /* 既存薬のn数*/, y      = /* 既存薬の有効数 */
                 ,Alpha  =  /* 有意水準    */, Delta = /* 非劣性マージン (0～1)*/
                 ,Method =  /* 手法の選択 (EXACT or ELIKE) */
                 ) ;
 DATA Noninf_EXL ;
```

```
keep N x M y Alpha Delta P1h P2h RiskDiff P1L P2L Pvalue Decison ;


*==== 初期設定 ====*;
N = &N ; x = &x ;       M = &M ; y = &y ;
Alpha = &Alpha ;        Delta = &Delta ;
pi = constant("pi") ;   p_intval = 0.001 ;

*==== 各群の比率・比率の差を算出 ====*;
P1h = x / N ; P2h = y / M ;
RiskDiff = P1h - P2h ;

*=== 二項比率の差の最尤推定を算出 P2を推定 ====*;
a = N + M ;
b = (-1) * ( N + M + x + y + Delta * ( N + 2 * M ) ) ;
c = M * ( Delta ** 2 ) + Delta * ( 2 * y + N + M ) + x + y ;
d = -1 * y * Delta * ( 1 + Delta ) ;
v = ( b / ( 3 * a ) ) ** 3 - b * c / ( 6 * ( a ** 2 ) ) + d / ( 2 * a ) ;
u = sign(v) * sqrt( b ** 2 / ( 3 * a ) ** 2 - c / ( 3 * a ) ) ;
* vが極小の場合の処理 *;
if v = 0 then w = ( Pi + arcos( 0 ) ) / 3 ;
else do ;
  wcos = v / ( u ** 3 ) ;
  w = ( pi + arcos( wcos ) ) / 3 ;
end ;
PL = 2 * u * cos( w ) - b / ( 3 * a ) ;

*==== 推定値から検定統計量を算出 ====*;
P1L = PL - Delta;
P2L = PL ;
Omega = sqrt( ( P1L  * ( 1 - P1L ) ) / N + ( P2L * ( 1 - P2L ) ) / M) ;
Z1 = ( P1h - P2h + Delta ) / Omega ;

%if %upcase( &Method ) = EXACT %then %do ;
*==== Exact Method ====*;
  * P-value の算出 *;
  * 各p = p2で、各xi,yiの組合せの尤度、検定統計量を算出、P-valueを計算 *;
  Pvalue = 0 ;
```

```
    sta_cnt = Delta / p_intval ;
    end_cnt = 1 / p_intval ;
    sta_cnt = round( sta_cnt , 0.1 ) ;  end_cnt = round( end_cnt , 0.1 ) ;
    do pcnt = sta_cnt to end_cnt by 1 ;
      P = pcnt * p_intval ;
      Pv_cnt = 0 ;
      do xi = 0 to N ;
        do yi = 0 to M ;
          nCx = comb( N , xi ) ;   mCy = comb( M , yi ) ;
          * 尤度の計算 0^0のための処理を追加 *;
          if ( P - Delta = 0 ) and ( xi = 0 ) then e = 1 ;
          else e = ( P - Delta ) ** xi ;
          if ( 1 - P + Delta = 0 ) and ( N - xi = 0 ) then f = 1 ;
          else f = ( 1 - P + Delta ) ** ( N - xi ) ;
          if ( P - Delta = 0 ) and ( yi = 0 ) then g = 1 ;
          else g = P ** yi ;
          if ( 1 - P = 0 ) and ( M - yi = 0 ) then h = 1 ;
          else h = ( 1 - P ) ** ( M - yi) ;
          Ph0 =  nCx * mCy * e * f * g * h ;
          Omega = sqrt( ( P  * ( 1 - P ) ) / N
                + ( ( P - Delta ) * ( 1 - P + Delta ) ) / M) ;
          P1i = xi / N ;  P2i = yi / M ;
          Z1i = ( P1i - P2i + Delta ) / Omega ;
          if Z1 <= Z1i then Pv_cnt = sum( Pv_cnt , Ph0 ) ;
        end ;
      end ;
      if Pvalue < Pv_cnt then Pvalue = Pv_cnt ;
    end ;

  %end ;

  %else %if %upcase( &Method ) = ELIKE %then %do ;
  *==== Exact Like Method ====*;
    * P = P2Lで、あるxi,yiのときの尤度、検定統計量を算出、P-valueを計算 *;
    P = P2L ;
    Pv_cnt = 0 ;
    do xi = 0 to N ;
      do yi = 0 to M ;
```

```
        nCx = comb( N , xi ) ;  mCy = comb( M , yi ) ;
        * 尤度の計算 0^0のための処理を追加 *;
        if ( P - Delta = 0 ) and ( xi = 0 ) then e = 1 ;
        else e = ( P - Delta ) ** xi ;
        if ( 1 - P + Delta = 0 ) and ( N - xi = 0 ) then f = 1 ;
        else f = ( 1 - P + Delta ) ** ( N - xi ) ;
        if ( P - Delta = 0 ) and ( yi = 0 ) then g = 1 ;
        else g = P ** yi ;
        h = ( 1 - P ) ** ( M - yi) ;
        Ph0 =  nCx * mCy * e * f * g * h ;
        Omega = sqrt( ( ( P - Delta ) * ( 1 - P + Delta ) ) / N
              + ( P * ( 1 - P ) ) / M) ;
        P1i = xi / N ;  P2i = yi / M ;
        Z1i = ( P1i - P2i + Delta ) / Omega ;
        if Z1 <= Z1i then Pv_cnt = sum( Pv_cnt , Ph0 ) ;
      end ;
    end ;
    Pvalue = Pv_cnt ;
  %end ;

  *==== P-value の切り上げ ====*;
  Pvalue = ceilz( Pvalue * 10000 ) / 10000 ;

  *==== 有意水準との判定 ====*;
  if Pvalue <= Alpha then Decison = "*" ;
  else Decison = "" ;

  output Noninf_EXL ;
RUN ;

*===Report Output====*;
PROC REPORT data = Noninf_EXL nowd ls = 150 ps = 30 center split="/" ;
  column ("InData Information /" ("Group1" N x P1h) ("Group2" M y P2h) ("RiskDiff" RiskDiff))
         ("Estimate / under the Null Hypothesis /" P1L P2L Delta)
         ("Non-Inferiority / Test Result /" Alpha Pvalue Decison)
  ;
  define N        / display width = 8  center "N" ;
  define x        / display width = 8  center "Sample" ;
```

```
  define P1h      / display width = 12 center "Proportion" ;
  define M        / display width = 8  center "N" ;
  define y        / display width = 8  center "Sample" ;
  define P2h      / display width = 12  center "Proportion" ;
  define RiskDiff / display width = 14 center "Group1-Group2" ;
  define P1L      / display width = 8  center "Group1" format = 8.4 ;
  define P2L      / display width = 8  center "Group2" format = 8.4 ;
  define Delta    / display width = 8  center "Delta" ;
  define Alpha    / display width = 8  center "Alpha" ;
  define Pvalue   / display width = 8  center "P-Value" format = 8.4 ;
  define Decison  / display width = 8  center "Decison" ;
RUN ;

%MEND Noninf_EXL ;
```

4.2. プログラム実行例

```
%Noninf_EXL(N=100, M=100, x=30, y=60, Alpha=0.05, Delta=0.4, Method=Exact) ;
```

InData Information												
Group1			Group2			RiskDiff	Estimate under the Hypothesis			Non-Inferiority Exact Test		
N	Sample	Proportion	N	Sample	Proportion	Group1-Group2	Group1	Group2	Delta	Alpha	P-Value	Decision
100	30	0.3	100	60	0.6	-0.3	0.2544	0.6544	0.4	0.05	0.0723	

```
%Noninf_EXL(N=100, M=100, x=30, y=60, Alpha=0.05, Delta=0.4, Method=elike) ;
```

InData Information												
Group1			Group2			RiskDiff	Estimate under the Hypothesis			Non-Inferiority elike Test		
N	Sample	Proportion	N	Sample	Proportion	Group1-Group2	Group1	Group2	Delta	Alpha	P-Value	Decision
100	30	0.3	100	60	0.6	-0.3	0.2544	0.6544	0.4	0.05	0.0676	

参考・引用文献

Farrington, C.P. and G. Manning (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*, 9, pp. 1447-1454.

Hauck, W. W. and Anderson, S. (1986), A comparison of large-sample confidence interval methods for the difference of two binomial probabilities, *The American Statistician*, 40, 318–322.

Kang, S. and J. J. Chen(2000). An approximate unconditional test of non-inferiority between two proportions. *Statistics In Medicine*, 19, pp. 2089-2100.

線形モデルにおける CLASSステートメントの機能

○吉田早織[1] 魚住龍史[2]
[1] 日本化薬株式会社 医薬データセンター
[2] 京都大学大学院医学研究科

The fascinating features for the CLASS in the context of linear models

Saori Yoshida[1] and Ryuji Uozumi[2]
[1] Clinical Data Management and Biostatistics, Nippon Kayaku Co., Ltd
[2] Kyoto University Graduate School of Medicine

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

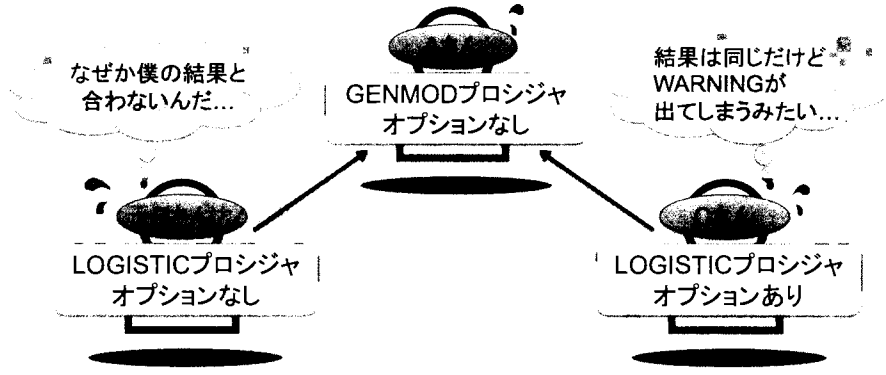
多くの線形モデルのプロシジャにサポートされているCLASSステートメントについて，デザイン行列（ダミー変数行列）の生成方法とプロシジャごとの デフォルトの整理，およびオプション機能について紹介する.

キーワード:CLASS, LOGISTIC, GENMOD, PARAM, ESTIMATE

社内でのある出来事

カテゴリカルな

課題 ある応答変数(2値データ)に対して，説明変数を用いて
モデル化し，各薬剤群における推定を行いたい



3

CLASSステートメントの役割

CLASSステートメント: 自動的にダミー変数を作成

自分での
設定が不要

多くのプロシ
ジャで使える

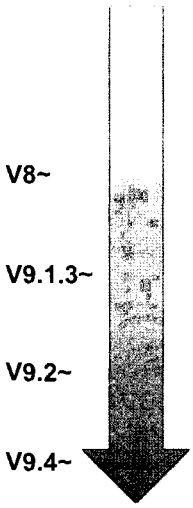
MODELステートメント: モデルの推定結果の出力

モデル推定後のプロセス: LSMEANS,
CONTRAST, ESTIMATE, LSMESTIMATE

**プロシジャごとにダミー変数生成方法の
デフォルト＆オプションが異なる**

4

線形モデルにおけるCLASSステートメント



一般線形モデル
➲ ~~REG~~, GLM

V8~

ロジスティックモデル
➲ LOGISTIC

浜田 (2000)

V9.1.3~ ➲ GLMSELECT

V9.2~

比例ハザードモデル
➲ PHREG

V9.4~ ➲ HPREG

5

今回のトピック

◆線形モデルにおけるダミー変数

◆LOGISTICプロシジャによるダミー変数の指定例

◆CLASSステートメントのWARNING&ERROR

◆3人 (Aさん, Bさん, Cさん) のSASプログラム及び出力結果の検証

6

線形モデルにおけるダミー変数 (1)

◆カテゴリーを数値(0,1)で表したもの

- ex) Treat がA,B,Pという変数を持つとき

$$\beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3$$

ダミー変数行列1

	x_1	x_2	x_3
A	1	0	0
B	0	1	0
P	0	0	1

- Treat Aの効果:

$$\beta_0 + \beta_1 \times 1 + \beta_2 \times 0 + \beta_3 \times 0 = \beta_0 + \beta_1$$

- Treat A と P の効果の差

$$\begin{array}{rl} \text{Treat A:} & \beta_0 + \beta_1 \\ \text{-) Treat P:} & \beta_0 + \beta_3 \\ \hline & \beta_1 - \beta_3 \end{array}$$

線形モデルにおけるダミー変数 (2)

◆カテゴリーを数値(0,1,-1)で表したもの

- ex) Treat がA,B,Pという変数を持つとき

$$\beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2$$

ダミー変数行列2

	x_1	x_2
A	1	0
B	0	1
P	-1	-1

- Treat Aの効果

$$\beta_0 + \beta_1 \times 1 + \beta_2 \times 0 = \beta_0 + \beta_1$$

- Treat A と P の効果の差

$$\begin{array}{rl} \text{Treat A:} & \beta_0 + \beta_1 \\ \text{-) Treat P:} & \beta_0 - \beta_1 - \beta_2 \\ \hline & 2\beta_1 + \beta_2 \end{array}$$

ダミー変数の生成例

効果法

PARAM=EFFECT

分類	値	デザイン変数	
A	A1	A2	A5
1	1	0	0
2	0	1	0
5	0	0	1
7	-1	-1	-1

GLM法

PARAM=GLM

分類	値	デザイン変数		
A	A1	A2	A5	A7
1	1	0	0	0
2	0	1	0	0
5	0	0	1	0
7	0	0	0	1

参照法

PARAM=REFRECT

分類	値	デザイン変数	
A	A1	A2	A5
1	1	0	0
2	0	1	0
5	0	0	1
7	0	0	0

累積順序法

PARAM=ORDINAL

分類	値	デザイン変数	
A	A1	A2	A5
1	0	0	0
2	1	0	0
5	1	1	0
7	1	1	1

多項式法

PARAM=POLYNOMIAL

分類	値	デザイン変数	
A	A1	A2	A5
1	1	1	1
2	2	4	8
5	5	25	125
7	7	49	343

9

LOGISTICプロシジャによる ダミー変数の指定

◆CLASSステートメントで PARAM= を指定することで ダミー変数の生成方法を指定できる

```
proc logistic data = Neuralgia ;
  class Treatment Sex /param=effect;
  model Pain = Treatment Sex ;
run;
```

＜SASデータにモデルを当てはめる＞

- SAS Helpに記載されているデータ "Neuralgia"
- Treatment(A,B,P), Sex(F,M)を説明変数, Pain (Yes,No)を応答変数としたモデル

10

効果法(PARAM=EFFECT)

◆平均効果との違いを比較

分類変数の水準の情報

分類	値	デザイン変数	
Treatment	A	1	0
	B	0	1
	P	-1	-1
Sex	F	1	
	M	-1	

最尤推定値の分析

パラメータ		自由度	推定値	標準誤差	Wald カイ 2 乗	Pr > ChiSq
Intercept		1	-0.4338	0.3224	1.8105	0.1785
Treatment	A	1	-0.8676	0.4623	3.5219	0.0606
	B	1	-0.8676	0.4623	3.5219	0.0606
Sex	F	1	-0.8959	0.3568	6.304	0.012

► Treatment A の推定

切片　Aを指定　平均を指定

```
estimate "estimate A" int 1 Treatment 1 0 Sex 0;
```

推定

ラベル	推定値	標準誤差	z 値	Pr > |z|
estimate A	-1.3013	0.5739	-2.27	0.0234

11

GLM法(PARAM=GLM)

◆最後のレベルとの比較

分類変数の水準の情報

分類	値	デザイン変数		
Treatment	A	1	0	0
	B	0	1	0
	P	0	0	1
Sex	F	1	0	
	M	0	1	

最尤推定値の分析

パラメータ		自由度	推定値	標準誤差	Wald カイ 2 乗	Pr > ChiSq
Intercept		1	2.1972	0.7566	8.4344	0.0037
Treatment	A	1	-2.6027	0.8434	9.5237	0.002
	B	1	-2.6027	0.8434	9.5237	0.002
	P	0	0	.	.	.
Sex	F	1	-1.7918	0.7136	6.304	0.012
	M	0	0	.	.	.

► Treatment A の推定

切片　Aを指定　平均を指定

```
estimate “estimate A” int 1 Treatment 1 0 0 Sex 0.5 0.5 ;
```

推定

ラベル	推定値	標準誤差	z 値	Pr > |z|
estimate A	-1.3013	0.5739	-2.27	0.0234

12

参照法(PARAM=REF)

◆参照レベルとの比較

分類変数の水準の情報

分類	値	デザイン変数	
Treatment	A	1	0
	B	0	1
	P	0	0
Sex	F	1	
	M	0	

最尤推定値の分析

パラメータ		自由度	推定値	標準誤差	Wald カイ 2 乗	Pr > ChiSq
Intercept		1	-2.1972	0.7566	8.4344	0.0037
Treatment	A	1	2.6027	0.8434	9.5237	0.002
	B	1	2.6027	0.8434	9.5237	0.002
Sex	F	1	1.7918	0.7136	6.304	0.012

Treatment A の推定

切片　Aを指定　平均を指定

```
estimate "estimate A" int 1 Treatment 1 0 Sex 0.5 ;
```

推定

ラベル	推定値	標準誤差	z 値	Pr > |z|
estimate A	-1.3013	0.5739	-2.27	0.0234

13

累積順序法(PARAM=ORDINAL)

◆順序のある効果間の比較

分類変数の水準の情報

分類	値	デザイン変数	
Treatment	P	0	0
	B	1	0
	A	1	1
Sex	F	0	
	M	1	

最尤推定値の分析

パラメータ		自由度	推定値	標準誤差	Wald カイ 2 乗	Pr > ChiSq
Intercept		1	0.4055	0.5839	0.4822	0.4874
Treatment	B	1	-2.6027	0.8434	9.5237	0.002
	A	1	5.50E-17	0.7785	0	1
Sex	F	1	1.7918	0.7136	6.304	0.012

Treatment A の推定

切片　Aを指定　平均を指定

```
estimate "estimate A" int 1 Treatment 1 1 Sex 0.5 ;
```

推定

ラベル	推定値	標準誤差	z 値	Pr > |z|
estimate A	-1.3013	0.5739	-2.27	0.0234

14

多項式法(PARAM=POLY)

◆2乗，3乗の効果を推定する

分類変数の水準の情報			
分類	値	デザイン変数	
Treatment	P	1.000	1.000
	B	2.000	4.000
	A	3.000	9.000
Sex	F	1.000	
	M	2.000	

最尤推定値の分析						
パラメータ		自由度	推定値	標準誤差	Wald カイ 2 乗	Pr > ChiSq
Intercept		1	3.8191	2.4135	2.504	0.1136
Treatment	POLY1	1	-6.5067	2.8429	5.2386	0.0221
	POLY2	1	1.3013	0.6934	3.5219	0.0606
Sex	POLY1	1	1.7918	0.7136	6.304	0.012

- Treatment A の推定

切片　Aを指定　平均を指定

```
estimate "estimate A" int 1 Treatment 3 9 Sex 1.5 ;
```

推定				
ラベル	推定値	標準誤差	z 値	Pr > |z|
estimate A	-1.3013	0.5739	-2.27	0.0234

15

CLASSステートメントの詳細設定

変数の順序設定

- ORDER＝DATA | FORMATED | FREQ | INTERNAL
  - DATA:データセットに出力した順
  - FORMATTED:FORMAT順(デフォルト)
  - FREQ:数が多い順
  - INTERNAL:フォーマットされていない値順
- DESCENDING
  - 降順にする

基準値の設定

- REF='label' | FIRST | LAST

PARAM=EFFECT
PARAM=REF
のときのみ

  - 'label'で基準変数値の選択
  - FIRST, LASTで最初の変数か最後の変数を選択

16

CLASSステートメントの WARNING＆ERROR(1)

Q REF=は，EFFECT法 or REF法でしか使えないが，他の手法で基準を指定するには？

ex) ordinal法

0	0
1	0
1	1

プラセボ基準にしたい

0	0
1	0
1	1

A Descendingで順番を変える

```
class TREATMENT / param=ordinal  descending ;
```

17

CLASSステートメントの WARNING＆ERROR(2)

Q LSMEANSステートメント，SLICEステートメントを使うとWARNINGが出力される

```
proc logistic data = data ;
  class Treatment Sex ;
  model Pain = Treatment Sex ;
  lsmeans  Treatment ;
run;
```

WARNING: The model does not have a GLM parameterization. This parameterization is required for the LSMEANS, LSMESTIMATE, and SLICE statement. These statements are ignored.

A LSMEANSステートメント，SLICEステートメントはGLM法のみ

```
class Treatment Sex / param=glm ;
```

18

CLASSステートメントの WARNING＆ERROR(3)

Q REF='label'でエラーが出力される

ex) TREATMENT＝'B' を基準としたい

```
class TREATMENT SEX / param=ref  ref = 'B' ;
```

ERROR 22-322: 構文エラーです。次の1つを指定してください: FIRST, LAST.

A 変数ごとに個別に設定する

Treatment	A	1	0
	B	0	1
	P	0	0
Sex	F	1	
	M	0	

➡

Treatment	A	1	0
	B	0	0
	P	0	1
Sex	F	0	
	M	1	

```
class TREATMENT(ref = 'B' )  SEX(descending) / param=ref ;
```

19

CLASSステートメントの WARNING＆ERROR(4)

Q GLIMMIXプロシジャでORDER=を用いるとエラーが出力される

```
proc glimmix data=data ;
  class Treatment Sex / order=first ;
  model Pain = Treatment Sex ;
run ;
```

ERROR 22-322: 構文エラーです。次の1つを指定してください: ;, TRUNCATE.
ERROR 202-322: オプションまたはパラメータを認識できません。無視します。

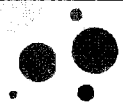
A ORDER=（順序）の指定は，PROCステートメント内で記述できる（詳細な指定はできない）

・GLM法のみのプロシジャはPROCステートメント内となる

```
proc glimmix data=data order=internal;
```

20

各プロシジャにおける ダミー変数生成法のサポート

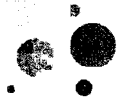


PARAM=	GLM	GLM SELECT	LOGISTIC	GEMNOD	PHREG	MIXED	GLIMMIX	FMM
EFFECT		○	☆	○	○			
GLM	☆	☆	○	☆	○	☆	☆	☆
ORDINAL THERMOMETER		○	○	○	○			
POLYNOMIAL POLY		○	○	○	○			
REFERENCE REF		○	○	○	☆			
ORTHEFFECT		○	○	○	○			
ORTHORDINAL ORTHOTHERM		○	○	○	○			
ORTHPOLY		○	○	○	○			
ORTHREF		○	○	○	○			

☆：デフォルト

21

3人のプログラムを検証



Aさん

```
proc genmod data=Neuralgia descending;
  class Treatment Sex ;
  model Pain = Treatment Sex  / link=logit dist=bin;
  estimate "estimate A" int 1 Treatment 1 0 0 Sex 0.5 0.5;
run;
```

対比推定の結果

ラベル	平均 推定値	平均 信頼限界		L'Beta 推定値	標準誤差	アルファ	L'Beta 信頼限界		カイ2乗 値	Pr > ChiSq
estimate A	0.2139	0.081	0.456	-1.3013	0.5739	0.05	-2.4261	-0.1766	5.14	0.0234

Bさん

```
proc logistic data=Neuralgia;
  class Treatment Sex ;
  model Pain(EVENT='Yes') = Treatment Sex;
  estimate “estimate A“
           int 1 Treatment 1 0 0 Sex 0.5 0.5;
run;
```

推定

ラベル	推定値	標準誤差	z 値	Pr > |z|
estimate A	-1.7493	0.647	-2.7	0.0069

Cさん

```
proc logistic data=Neuralgia;
  class Treatment Sex / param=ref;
  model Pain(EVENT='Yes') = Treatment Sex;
  estimate "estimate A"
           int 1 Treatment 1 0 0 Sex 0.5 0.5;
run;
```

推定

ラベル	推定値	標準誤差	z 値	Pr > |z|
estimate A	-1.3013	0.574	-2.27	0.0234

22

Aさんの方法

GENMODプロシジャ オプションなし

```
proc genmod data=Neuralgia descending;
  class Treatment Sex ;
  model Pain = Treatment Sex  / link=logit dist=bin;
  estimate "estimate A" int 1 Treatment 1 0 0 Sex 0.5 0.5;
run;
```

GENMODプロシジャの
デフォルトはGLM法

◆ Treatment Aの効果を見るため他の効果（Sex）は平均とする

$$y_A = \beta_0 + \beta_{Treat1} + 0.5\beta_{Sex1} + 0.5\beta_{Sex2}$$

GLM法		x_1	x_2	x_3
Treatment	A	1	0	0
	B	0	1	0
	P	0	0	1
Sex	F	1	0	
	M	0	1	

Bさんの方法

LOGISTICプロシジャ オプションなし

```
proc logistic data=Neuralgia;
  class Treatment Sex ;
  model Pain(EVENT='Yes') = Treatment Sex;
  estimate “estimate A“  int 1 Treatment 1 0 0 Sex 0.5 0.5;
run;
```

WARNING: More coefficients than levels specified for effect Treatment. Some coefficients will be ignored.

LOGISTICプロシジャの
デフォルトはEFFECT法

◆ Treatment Aの効果を見るため他の効果（Sex）は平均とする（0）

$$y_A = \beta_0 + \beta_{Treat1}$$

参 estimate “estimate A“ int 1 Treatment 1 ;

EFFECT法		x_1	x_2
Treatment	A	1	0
	B	0	1
	P	-1	-1
Sex	F	1	
	M	-1	

Cさんの方法

LOGISTICプロシジャ PARAMオプションあり

```
proc logistic data=Neuralgia;
  class Treatment Sex / param=ref;
  model Pain(EVENT='Yes') = Treatment Sex;
  estimate "estimate A"  int 1 Treatment 1 0 0 Sex 0.5 0.5;
run;
```

WARNING: More coefficients than levels specified for effect Treatment. Some coefficients will be ignored.

REF法を指定

REF法		x_1	x_2
Treatment	A	1	0
	B	0	1
	P	0	0
Sex	F	1	
	M	0	

◆ Treatment Aの効果を見るため他の効果（Sex）は平均とする（0）

$$y_A = \beta_0 + \beta_{Treat1} + 0.5\beta_{Sex}$$

参 estimate "estimate A" int 1 Treatment 1 0 Sex 0.5 ;

25

まとめ

カテゴリカルな

ある応答変数（2値データ）に対して，説明変数を用いてモデル化し，<u>各薬剤群における推定</u>を行いたい

GLM法		x_1	x_2	x_3
Treatment	A	1	0	0
	B	0	1	0
	P	0	0	1
Sex	F	1	0	
	M	0	1	

LSMEANS, CONTRAST, ESTIMATE, LSMESTIMATE

26

参考文献

1. Carpenter AL. Programming With CLASS: Keeping Your Options Open. Proceedings of the SAS Global Forum. Cary, NC: SAS Institute Inc., 2014. Available at http://support.sas.com/resources/papers/proceedings14/1270-2014.pdf.
2. Pasta DJ. Parameterizing Models to Test the Hypotheses You Want: Coding Indicator Variables and Modified Continuous Variables. Proceedings of the 30th Annual SAS Users Group International Conference. Cary, NC: SAS Institute Inc., 2005. Available at http://www2.sas.com/proceedings/sugi30/212-30.pdf.
3. Pritchard ML, Pasta, DJ. Head of the CLASS: Impress your colleagues with a superior understanding of the CLASS statement in PROC LOGISTIC. Proceedings of the 29th Annual SAS Users Group International Conference. Cary, NC: SAS Institute Inc., 2004. Available at http://www2.sas.com/proceedings/sugi29/194-29.pdf.
4. SAS Institute Inc. SAS/STAT(R) 9.3: User's Guide. Cary, NC, USA: SAS Institute Inc., 2011.
5. SAS Institute Inc. SAS/STAT(R) 12.3: User's Guide: High-Performance Procedures. Cary, NC, USA: SAS Institute Inc., 2013.
6. 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.
7. 竹内啓, 高橋行雄, 大橋靖雄, 芳賀敏郎. SASによる実験データの解析. 東京大学出版会, 1989.
8. 浜田知久馬. V.8におけるLOGISTICの機能拡張. 日本SASユーザー会(SUGI-J) 論文集 2000, 13–38.
9. 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, 185–199.

SASによる二項比率における正確な信頼区間の比較

原茂恵美子[1)] 武藤彬正[1)] 宮島育哉[2)] 榊原伊織[2)]

1) 株式会社タクミインフォメーションテクノロジー システム開発推進部

2) 株式会社タクミインフォメーションテクノロジー ビジネスソリューション部

Comparison of Five Exact Confidence Intervals for the Binomial Proportion using SAS

Emiko Haramo[1)] Akimasa Muto[1)] Ikuya Miyajima[2)] Iori Sakakibara[2)]

1) System Development Department, Takumi Information Technology Inc.

2) Business Solutions Department, Takumi Information Technology Inc.

Abstract

SAS has Clopper–Pearson confidence interval based on an exact test using binominal proportions. But it has been indicated that this method is extremely conservative. This study introduces five exact confidence intervals where the actual coverage probability does not fall under the nominal coverage probability. These confidence intervals contain the methods not implemented in SAS. Further, we calculate the expected length of the confidence intervals and compare/verify the accuracy of the coverage probabilities. As a result, we found that the quality of the confidence interval based on the Sterne test is its availability for small samples. We also present a SAS macro program about a calculation of confidence interval.

Key words: Binominal Proportion; Confidence Interval; Exact Method

1. Introduction

Studies on confidence intervals for binomial proportions have been performed since a long time ago and continue to be performed. Because the Wald interval is easy to calculate, it is often used as the confidence interval for binomial proportions. However, when using this confidence interval, the actual coverage probability often falls under the nominal coverage probability in small cases.

On the other hand, several confidence intervals where the actual coverage probability does not fall under the nominal coverage probability are suggested. Clopper and Pearson (1934) suggest a construction method for a confidence interval based on an exact test using binominal proportions. The confidence interval is a method for the actual coverage probability not to fall under the nominal coverage probability at all times, but it has been indicated that this method is extremely conservative (Agresti and Coull (1998)). In addition, several other exact methods have been suggested. However, no papers have compared these exact confidence intervals in detail.

This study introduces five exact confidence intervals where the actual coverage probability does not fall under the nominal coverage probability; moreover, we calculate the expected length of the confidence interval and compare/verify the accuracy of the coverage probabilities.

2. Notation and Method

Let X be independent random variables. Suppose that X follows a binominal distribution with parameters n, π.

2.1 Clopper–Pearson Confidence Interval

The Clopper–Pearson confidence interval can be written as

$$\left[\frac{x}{x+(n-x+1)F_{2(n-x+1),2x}(\alpha/2)}, \frac{x+1}{x+1+(n-x)/F_{2(x+1),2(n-x)}(\alpha/2)}\right]$$

where $F_{a,b}(\alpha)$ is the upper $100(\alpha/2)\%$ quartile from an F-distribution with a and b degrees of freedom.

2.2 Exact Likelihood Ratio Confidence Interval

The exact likelihood ratio (LR) confidence interval is based on inverting the acceptance regions for the exact binomial tests of H_0: $\pi = \pi_0$. Following Fleiss *et al.* (2003) and given α and true $\pi = \pi_0$, we define the generalized log LR (GLLR) statistic as

$$\text{GLLR}(\pi_0 \mid x) = x \ln\{x/(n\pi_0)\} + (n - x) \ln\{(n - x)/\{n(1 - \pi_0)\}$$

For $\pi_0 = 0$, GLLR is only defined for $x = 0$; for $\pi_0 = 1$, GLLR is only defined for $x = n$. We define the attained LR p-value as

$$p-\text{value} = \sum_t \binom{n}{t} \pi_0(1-\pi_0),$$

where the sum is taken over the set t of x_i values for which $\text{GLLR}(\pi_0 \mid x_i) \geq \text{GLLR}(\pi_0 \mid x)$, excluding those values where GLLR is not defined. Then, the exact LR confidence set is the set of all π_0 such that the p-value $\geq \alpha$.

2.3 Exact Score Confidence Interval

The exact score confidence (SC) interval is based on inverting the acceptance regions for the exact score tests of H_0: $\pi = \pi_0$. Following Hirji (2006) and given α and true $\pi = \pi_0$, we define the score statistic as

$$\text{SC}(\pi_0 \mid x) = (x - n\pi_0)^2 / (n\pi_0 (1 - \pi_0))$$

For $\pi_0 = 0$, SC is only defined for $x = 0$ and SC = 0. For $\pi_0 = 1$, SC is only defined for $x = n$ and SC = 0. We define the attained score p-value as

$$p-\text{value} = \sum_t \binom{n}{t} \pi_0(1-\pi_0)$$

where the sum is taken over the set t of x_i values for which $\text{SC}(\pi_0 \mid x_i) \geq \text{SC}(\pi_0 \mid x)$, excluding those values where SC is not defined.Then, the exact score confidence set is the set of all π_0 such that the p-value $\geq \alpha$.

2.4 Sterne Confidence Interval

The interval proposed by Reiczigel (2003) is defined by inverting the exact binomial test with acceptance regions, including the most probable values of the binomial variable, and then taking the most probable, followed by the next most probable, until their total probability reaches the required level, for example, 95%.

Assume that we want to invert a test of H_0: $\pi_1 = \pi_0$ for the binomial parameter π to obtain a 95% confidence interval for π based on $n = 5$ observations. Denote X_1 to be the observed number of successes. The basic idea is that a 95% confidence set should consist of all such values π_0 of the parameter for which H_0: $\pi_1 = \pi_0$ is not rejected by the test at the 95% level. For simplicity, assume that a one-digit precision is sufficient for the interval endpoints, because in such a case, the procedure can be demonstrated using a small table of binomial probabilities (Table 1). In the case of $X_1 = 3$, the region 0.2 to 0.9 has become the acceptance region (see the underlined portion in Table 1); the minimum π is the lower confidence bound, and the maximum π is the upper confidence bound.

Table 1: Binomial probabilities

($n = 6, \alpha = 0.1$)

X_1	π										
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
6	0.0000	0.0000	0.0001	0.0007	0.0041	0.0156	0.0467	**0.1176**	**0.2621**	**0.5314**	**1.0000**
5	0.0000	0.0001	0.0015	0.0102	0.0369	**0.0938**	**0.1866**	**0.3025**	**0.3932**	**0.3543**	0.0000
4	0.0000	0.0012	0.0154	0.0595	**0.1382**	**0.2344**	**0.3110**	**0.3241**	**0.2458**	**0.0984**	0.0000
3	0.0000	0.0146	0.0819	**0.1852**	**0.2765**	**0.3125**	**0.2765**	**0.1852**	0.0819	0.0146	0.0000
2	0.0000	**0.0984**	**0.2458**	**0.3241**	**0.3110**	**0.2344**	**0.1382**	0.0595	0.0154	0.0012	0.0000
1	0.0000	**0.3543**	**0.3932**	**0.3025**	**0.1866**	**0.0938**	0.0369	0.0102	0.0015	0.0001	0.0000
0	**1.0000**	**0.5314**	**0.2621**	**0.1176**	0.0467	0.0156	0.0041	0.0007	0.0001	0.0000	0.0000

2.5 Blaker Confidence Interval

Blaker (2000) has proposed a new exact interval that is an excellent alternative to the Sterne interval, and that has many commonalities with the Sterne interval. The approach builds on the concept of acceptability function (ACC) (Blaker (2000)).

$$ACC(x) = \min\left\{\sum_{i=0}^{x} f(i,p), \sum_{j=x}^{n} f(j,p)\right\}$$

Table 2: The acceptance regions of Binomial probabilities

($n = 6, \alpha = 0.1$)

X	P_i										
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
6	-0.1000	-0.1000	-0.0999	-0.0993	-0.0959	-0.0844	-0.0124	**0.0881**	**0.2610**	**0.9000**	**0.9000**
5	-0.1000	-0.0999	-0.0984	-0.0891	-0.0590	**0.0250**	**0.3125**	**0.5759**	**0.9000**	**0.3686**	-0.1000
4	-0.1000	-0.0987	-0.0830	-0.0295	**0.1259**	**0.3531**	**0.9000**	**0.9000**	**0.5068**	**0.0143**	-0.1000
3	-0.1000	-0.0841	-0.0011	**0.2733**	**0.5890**	**0.9000**	**0.5890**	**0.2733**	-0.0011	-0.0842	-0.1000
2	-0.1000	**0.0143**	**0.5068**	**0.9000**	**0.9000**	**0.5875**	**0.1259**	-0.0295	-0.0830	-0.0987	-0.1000
1	-0.1000	**0.3686**	**0.9000**	**0.5759**	**0.3125**	**0.1188**	-0.0590	-0.0891	-0.0984	-0.0999	-0.1000
0	**0.9000**	**0.9000**	**0.2610**	**0.0881**	-0.0124	-0.0688	-0.0959	-0.0993	-0.0999	-0.1000	-0.1000
											ACC(x)-α

In the Appendix, we present a SAS macro program about a calculation method for the five confidence intervals.

3. Comparison of the five confidence intervals

In this study, the coverage probability and the expected length were used as the basis for our evaluation, and 95% of

each confidence interval was compared.

The coverage probabilities were computed using the proportion with which the confidence interval includes the binominal proportion. A simulation of 100,000 rounds of under defined values of π was conducted. Similarly, the expected lengths were computed using the mean of the difference in the confidence intervals.

3.1 Comparison of coverage probability

Figure 1 shows the coverage probabilities of the five exact confidence intervals. Overall, all methods described a high coverage probability for $\pi = 0$ and 1; in addition, the values were slightly higher near $\pi = 0.5$. This figure indicates that Clopper–Pearson is clearly a conservative method compared with the other methods. The results of the Exact GLLR and the Exact Score showed higher values depending on the value of π. In addition, the coverage probability of Sterne and Blaker were significantly closer to 0.95 than the other confidence intervals.

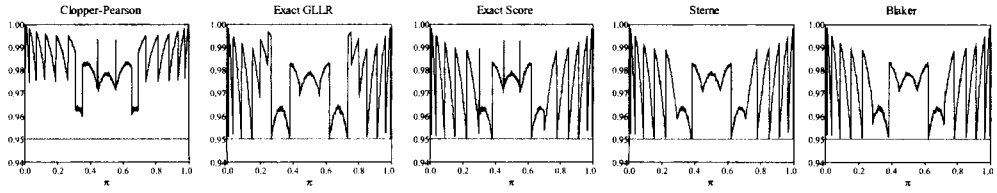


Figure 1 Coverage probabilities of the five exact confidence intervals
($n = 10$, a significance level of 0.05, and $0.001 \leq \pi \leq 0.999$)

The coverage probability appears close to 95% as n increases (Figure 2). The result of figure indicates that the coverage probability varies for Clopper–Pearson, the Exact Score, and Blaker by the value of n. For exact GLLR and Sterne intervals, the value variation is small and it is closer to 95% as n increases.

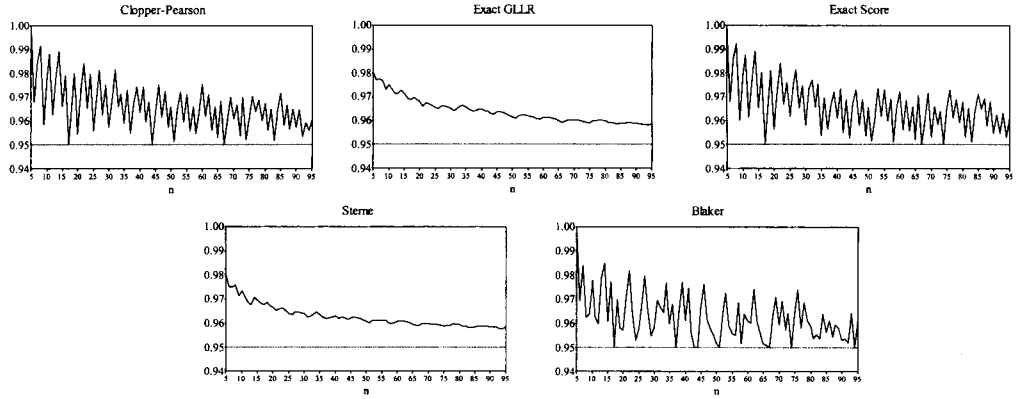


Figure 2 Coverage probabilities for $n = 5$ to 95
(a significance level of 0.05, and $\pi = 0.50$)

3.2 Comparison of Expected Length

Clopper–Pearson is clearly conservative compared with the other methods (Figure 3). For $\pi = 0$, the expected length values are smaller for GLLR; however, for $\pi = 0.5$, the values are larger. For the Score method, the values are smaller compared with the other methods when $\pi = 0.5$; however, for $\pi = 0$, the values are larger and varied. In addition, Sterne and Blaker are not scattered in the width compared with the other methods, and their values are similar.

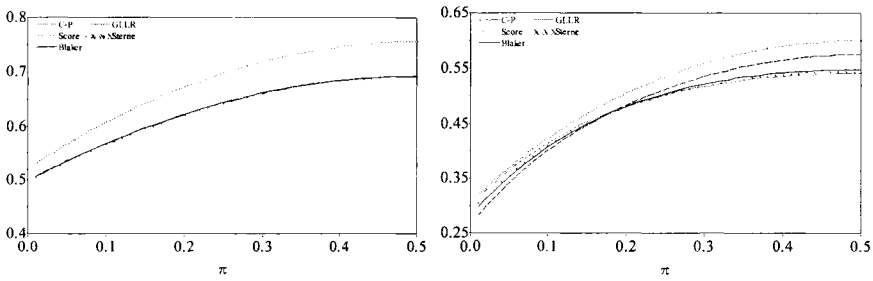


Figure 3 Expected Length

(Left: $n = 5$ and $0.001 \leq \pi \leq 0.999$, Right: $n = 10$ and $0.001 \leq \pi \leq 0.999$)

4. Conclusion

In this study, we introduce five exact confidence intervals. Exact GLLR, Exact Score, Sterne confidence interval and Blaker confidence interval are not implemented in SAS. Moreover, we proposed the program for the methods to calculate the confidence intervals of the binomial proportion. We examined five exact confidence intervals that do not fall under the nominal coverage probability in order to determine the most useful method for small sample sizes. We calculated the expected length of the confidence intervals and compared/verified the accuracy of the coverage probabilities. For the Clopper–Pearson method implemented in SAS, we found that the expected length and coverage probability is even more conservative than the other methods as a result of simulation of 100,000. For the exact GLLR method, the evaluated values were near the edge of the expected length; however, for $\pi = 0.5$, the values were conservative, similar to the Clopper–Pearson method. For the Exact Score method, the values in the coverage probability for $n = 10$ tended to appear high depending on π, and the variation was significantly related to variations of n. Moreover, the calculated expected length appeared scattered. The coverage probability of Sterne and Blaker were significantly closer to 0.95 than the other confidence intervals and for the expected lengths. However, the Blaker method showed scattering values of the coverage probabilities related to the variance of n, whereas the Sterne method was stable.

In summary, we considered the Sterne confidence interval method to be more useful than the other methods in small sample sizes.

5. References

Agresti, A. and Coull, B.A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician* 52, 119-126.

Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions. *The Canadian Journal of Statistics* 28, 783-798.

Clopper, C.J. and Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26, 404-413

Fleiss, J.L., Levin, B. and Paik, M.C. (2003). *Statistical methods for rates and proportions*, 3rd edition. New York: JohnWiley and Sons.

Hirji, K. (2006). Exact analysis of discrete data. New York: Chapman and Hall/CRC.

Reiczigel, J. (2003). Confidence intervals for the binomial parameter: some new considerations. *Statistics in Medicine* 22, 611-621.

Appendix

```
/*
+-------------------------------+-------------------------------------------------------------------------------+
| DESCRIPTION:                  | Takumi Information Technology
| VERSION:                      | SAS 9.4
| LANGUAGE:                     | Japanese
| PRODUCT:                      | BASE/GRAPH/STAT
| FILENAME:                     | Five_Exact.sas
+-------------------------------+-------------------------------------------------------------------------------+
| HISTORY:                      | E.Haramo              2014/05/29 Initial Coding
+-------------------------------+-------------------------------------------------------------------------------+
*/
options nodate ;

%macro Five_Exact ( alpha = , n = , x = ) ;

%* 1. Clopper-Pearson Confidence Interval *;
 data work.Out1 ;
    i = 0 ; val = &n - &x ; output ;
    i = 1 ; val =       &x ; output ;
 run ;

 proc freq data = work.Out1 noprint ;
    weight val / zeros ;
    table i / binomial alpha = &alpha;
    output out = work.Out2 binomial ;
 run ;

 data work.C_P ;
    set work.Out2 ;
    length STAT $ 100 ;
    STAT = "Clopper-Pearson" ;
    Low   = round( XL_BIN , 0.00001 ) ;
    Up    = round( XU_BIN , 0.00001 ) ;
    keep STAT Low Up ;
 run ;
```

```
%* 2. Exact GLLR Confidence Interval *;
 data work.Out1 ( drop = p_ ) ;
   length STAT $ 100 ;
   STAT = "GLLR" ;
   do p_ = 0 to 1000 by 1 ;
     p = p_ / 1000 ;
     if       &x =   0 and 0 <= P <   1
    then GLLR = ( ( &n - &x ) * log( ( &n - &x )
                / ( &n * ( 1 - p ) ) ) ) ;
    else if   &x = &n and 0 <   P <= 1
    then GLLR = ( &x * log( &x / ( &n * p ) ) ) ;
    else if                0 < P < 1
    then GLLR = ( &x * log( &x / ( &n * p ) ) )
              + ( ( &n - &x ) * log( ( &n - &x )
              / ( &n * ( 1 - p ) ) ) ) ;
    else GLLR = . ;
    do t = 0 to &n ;
      if       t =   0 and 0 <= P <   1
      then GLLRi = ( ( &n - t ) * log( ( &n - t )
                 / ( &n * ( 1 - p ) ) ) ) ;
      else if   t = &n and 0 <   P <= 1
      then GLLRi = ( t * log( t / ( &n * p ) ) ) ;
      else if                0 < P < 1
      then GLLRi = ( t * log( t / ( &n * p ) ) )
                 + ( ( &n - t ) * log( ( &n - t )
                 / ( &n * ( 1 - p ) ) ) ) ;
      else GLLRi = . ;
      if   t = 0
      then Bin = probbnml( p , &n , t ) ;
      else
      Bin = probbnml( p , &n , t )
          - probbnml( p , &n , t - 1 ) ;
      if n( GLLR , GLLRi ) = 2 and GLLRi >= GLLR
      then output ;
    end ;
   end ;
run ;

proc means data = work.Out1 nway noprint ;
   class STAT p ;
   var Bin ;
   output out = work.Out2
     ( where = ( pval >= &alpha ) ) sum = pval ;
run ;

proc means data = work.Out2 nway noprint ;
   class STAT ;
   var p ;
   output out = work.GLLR MIN = Low MAX = Up ;
run ;

%* 3. Exact Score Confidence Interval *;
 data work.Out1 ( drop = p_ ) ;
   length STAT $ 100 ;
   STAT = "Score" ;
   do p_ = 0 to 1000 by 1 ;
     p = p_ / 1000 ;
     if &x = 0 and P = 0 then SCORE = 0 ;
     else if 0 =< &x =< &n and 0 < P < 1
     then SCORE = ( ( &x - ( &n * p ) ) ** 2 )
                / ( &n * p * ( 1 - p ) ) ;
     else if &x = &n and P = 1 then SCORE = 0 ;
     else SCORE = . ;
     do t = 0 to &n ;
       if t = 0 and P = 0 then SCOREi = 0 ;
       else if 0 =< t =< &n and 0 < P < 1
       then SCOREi = ( ( t - ( &n * p ) ) ** 2 )
                   / ( &n * p * ( 1 - p ) ) ;
       else if t = &n and P = 1 then SCOREi = 0 ;
       else SCOREi = . ;
       if t = 0 then Bin = probbnml( p , &n , t ) ;
       else          Bin = probbnml( p , &n , t )
                         - probbnml( p , &n , t - 1 ) ;
       if n( SCORE , SCOREi ) = 2
          and SCOREi >= SCORE then output ;
```

```
      end ;
    end ;
run ;

proc means data = work.Out1 nway noprint ;
   class STAT p ;
   var Bin ;
   output out = work.Out2
     ( where = ( pval >= &alpha ) ) sum = pval ;
run ;

proc means data = work.Out2 nway noprint ;
   class STAT ;
   var p ;
   output out = work.Score MIN = Low MAX = Up ;
run ;

%* 4. Sterne Confidence Interval *;
data work.Out1;
   length STAT $ 100 ;
   STAT = "Sterne" ;
   do x = 0 to &n ;
     do p_ = 0 to 1000 ;
       p = p_ / 1000 ;
       Prob = pdf( 'binom' , x , p , &n ) ;
       output ;
     end ;
   end ;
run ;


proc sort data = work.Out1 ;
   by p descending Prob ;
run ;

data work.Out2 ;
   set work.Out1 ;
   by p ;
   if first.p then do ;
     Prob_AD = 0 ;
     BProb_AD = . ;
   FLG = 0 ;
   end ;
   Prob_AD + prob ;
   BProb_AD = lag1( Prob_AD ) ;
   if  ( BProb_AD < (1 - &alpha)) then FLG = 0 ;
   else                               FLG = 1 ;
run ;

proc means data = work.Out2 nway noprint ;
   where x = &x and FLG = 0 ;
   class STAT ;
   var p ;
   output out = work.Sterne MIN = Low MAX = Up ;
run ;

%* 5. Blaker Confidence Interval *;
data work.Out1 ;
   length STAT $ 100 ;
   STAT = "Blaker" ;
   do p_ = 0 to 1000 ;
     do x = 0 to &n ;
       p = p_ / 1000 ;
       if  x = 0 then p1 = 1 ;
       else    p1 = 1 - probbnml( p , &n , x - 1 ) ;
       p2 = probbnml( p , &n , x ) ;
       output ;
     end ;
   end ;
run ;

data work.Out2 ;
   set work.Out1;
   do u = &n to 0 by -1.0 ;
     px1 = probbnml( p , &n , u ) ;
     if px1 >=   p1       then x1 = u ;
     if px1 >= ( 1 - p2   ) then x2 = u ;
   end ;
```

```
    if     x1 = 0 then a1 = p1 ;
    else   a1 = p1 + probbnml( p , &n , x1 - 1 ) ;
    a2 = p2 + 1 - probbnml( p , &n , x2 ) ;
    accept = min( a1 , a2 ) ;
    if ( accept - &alpha ) >= 0 ;
  run ;

  proc means data = work.Out2 nway noprint ;
    where x = &x ;
    class STAT ;
    var p ;
    output out = work.Blaker MIN = Low MAX = Up ;
  run ;

Title "Exact Confidence Interval ( N = &N, X = &x )" ;

data work.OUTDS ;
  set work.C_P
      work.GLLR
      work.Score
      work.Sterne
      work.Blaker ;
run ;

proc print label noobs ;
  var STAT Low Up ;
  format Low Up 8.3 ;
  label STAT = " Exact Confidence Interval "
        Low = "Lower Confidence Interval "
        Up = "Upper Confidence Interval" ;
run ;
%mend ;

%Five_Exact( alpha = 0.05 , n = 10 , x = 3 ) ;
```

Exact Confidence Interval (N = 10, X = 3)

Exact Confidence Interval	Lower Confidence Interval	Upper Confidence Interval
Clopper-Pearson	0.393	0.913
GLLR	0.116	0.616
Score	0.116	0.600
Sterne	0.116	0.602
Blaker	0.116	0.606

オッズ比の信頼区間の構成法の比較

飯塚 政人, 猪嶋 恭平, 浜田 知久馬
東京理科大学 工学研究科

Comparison of confidence intervals for the odds ratio

Masato Iizuka , Kyohei Inoshima, Chikuma Hamada
Graduate School of Engineering,
Tokyo University of Science

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

LOGISTICプロシジャで利用可能なオッズ比の信頼区間の構成法を紹介する.
また, 各信頼区間の被覆確率等を比較した結果を示し, 推奨すべき方法を報告する.

キーワード: 信頼区間, オッズ比, 被覆確率, LOGISTICプロシジャ

発表構成

- 研究背景
- 研究目的
- 研究方法
- 結果・考察
- まとめと今後の課題

オッズ比について[2]

オッズ
イベントが起こる確率と起こらない確率の比

オッズ比
ある条件のもとのオッズと別の条件のもとのオッズの比

オッズ比は医薬研究でよく用いられる

標準的に用いられる信頼区間はWald信頼区間[3, 6]

2×2分割表

	群1	群2
陽性	n_{11}	n_{12}
陰性	n_{21}	n_{22}
計	$n_{\cdot 1}$	$n_{\cdot 2}$
母数	π_1	π_2

オッズ比(真値)

$$OR = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

オッズ比(推定値)

$$\widehat{OR} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

オッズ比の信頼区間

- Wald信頼区間
- WaldFirth信頼区間
- PL(尤度比)信頼区間
- PLFirth(尤度比Firth)信頼区間
- Exact(正確検定)信頼区間
- Midp信頼区間

LOGISTICプロシジャで使用可能な信頼区間

6

信頼区間の構成(1)

■ Wald信頼区間[2]

$$\exp\left(\log\widehat{OR} \pm \left(Z_{\alpha/2} \times \sqrt{Var\left(\log\widehat{OR}\right)}\right)\right)$$

- 分散: $Var\left(\log\widehat{OR}\right) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$
- $Z_{\alpha/2}$: 標準正規分布の$100(1-\frac{\alpha}{2})$%点

信頼区間の構成(2)

■ PL(尤度比)信頼区間[2]

$$-2\left(\log L(\theta_0) - \log L\left(\widehat{OR}\right)\right) < X_1^2(\alpha)$$

- $\log L$: 対数尤度
- $X_1^2(\alpha)$: カイ二乗分布の上側α%点
- θ_0: 信頼限界値

■ Firthの補正

■ 尤度に情報行列の行列式の0.5乗をかけ最尤推定

各セル度数に0.5を足すことに対応

信頼区間の構成(3)

■ Exact(正確検定)[2,3]

上限ω_U:$\sum_{x=\max(0,l)}^{x_0} H(x|n_{..}, n_{1.}, n_{.1}, \omega_U) = \frac{\alpha}{2}$

下限ω_L:$\sum_{x=x_0}^{\min(n_{1.},n_{.1})} H(x|n_{..}, n_{1.}, n_{.1}, \omega_L) = \frac{\alpha}{2}$

- 上側確率:$\sum_{x=x_0}^{\min(n_{1.},n_{.1})} H(x|n_{..}, n_{1.}, n_{.1}, \omega_0)$
- 下側確率:$\sum_{x=\max(0,l)}^{x_0} H(x|n_{..}, n_{1.}, n_{.1}, \omega_0)$

H:非心超幾何分布の確率関数

$l = n_{1.} + n_{.1} - n_{..}$

信頼区間の構成(4)

■ Midp信頼区間

上限ω_U:

$\sum_{x=\max(0,l)}^{x_0+1} H(x|n_{..}, n_{1.}, n_{.1}, \omega_U) + \frac{1}{2}H(X_0) = \frac{\alpha}{2}$

下限ω_L:

$\sum_{x=x_0+1}^{\min(n_{1.},n_{.1})} H(x|n_{..}, n_{1.}, n_{.1}, \omega_L) + \frac{1}{2}H(X_0) = \frac{\alpha}{2}$

2×2分割表の実例[5]

幼児の発達障害の結果を示す2×2分割表

	PVL群(ケース群)	コントロール群
有り	2	1
無し	24	25
合計	26	26

オッズ比の推定値: $\widehat{OR} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{2 \times 25}{1 \times 24} = 2.083$

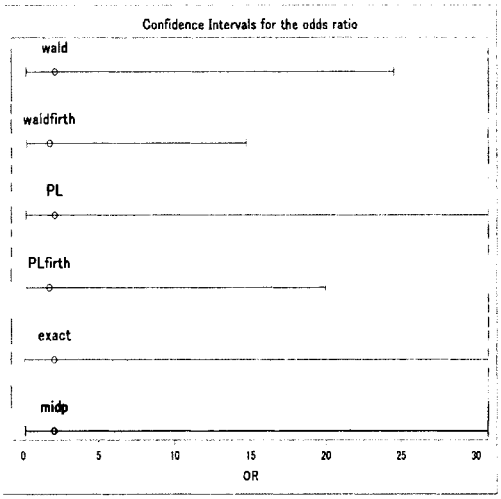
LOGISTICプロシジャで6つの信頼区間を出力

```
PROC LOGISTIC DATA = DATA;
    MODEL EXPOSURE = RESPONSE/
      CLODDS = BOTH;
    FREQ COUNT;
    EXACT RESPONSE /
      CLTYPE = EXACT  ESTIMATE = ODDS;
RUN;
```

Wald信頼区間, PL(尤度比)信頼区間を出力
CLODDS=BOTH FIRTH
とすることでFirth法を加味して出力

Exact(正確検定)信頼区間を出力
CLTYPE = MIDP ESTIMATE = ODDS;
とすることで, Midp信頼区間を出力

6種類の信頼区間の比較



信頼区間	信頼下限	信頼上限
Wald	0.177	24.51
WaldFirth	0.204	14.73
PL	0.188	46.53
PLFirth	0.215	19.96
Exact	0.101	127.76
Midp	0.148	63.75

※95%信頼区間算出

本研究の目的

➢症例数やオッズ比の真の値を変えて、6種類の手法について信頼区間の被覆確率を評価し、適切な方法を検討

信頼区間の被覆確率

□被覆確率(Coverage Probability) : $C(\Delta)$
➢信頼区間が真値を含む確率

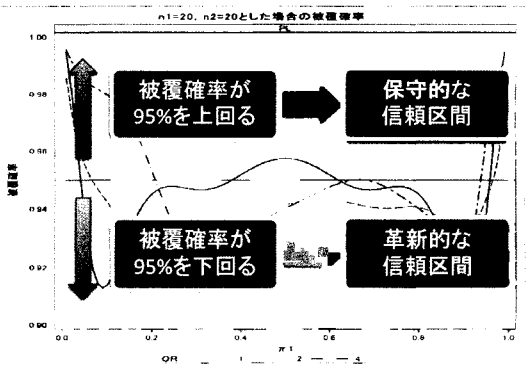
$$C(\pi_1,\pi_2)=\sum_{n_{11}=0}^{n_{1\cdot}}\sum_{n_{21}=0}^{n_{2\cdot}} I(n_{11},n_{21},\pi_1,\pi_2)$$
$$\times\binom{n_{1\cdot}}{n_{11}}{\pi_1}^{n_{11}}(1-\pi_1)^{n_{1\cdot}-n_{11}}\binom{n_{2\cdot}}{n_{21}}{\pi_2}^{n_{21}}(1-\pi_2)^{n_{2\cdot}-n_{21}}$$

$$I(n_{11},n_{12},\pi_1,\pi_2)=\begin{cases}1,\text{信頼区間が真値を含む}\\0,\text{信頼区間が真値を含まない}\end{cases}$$

信頼水準95%の信頼区間の場合,
被覆確率が95%に近い信頼区間が望ましい

信頼区間の評価基準

□2種類の評価基準で適切な信頼区間を検討
➢被覆確率が95%に最も近い信頼区間
➢被覆確率が保守的になる信頼区間

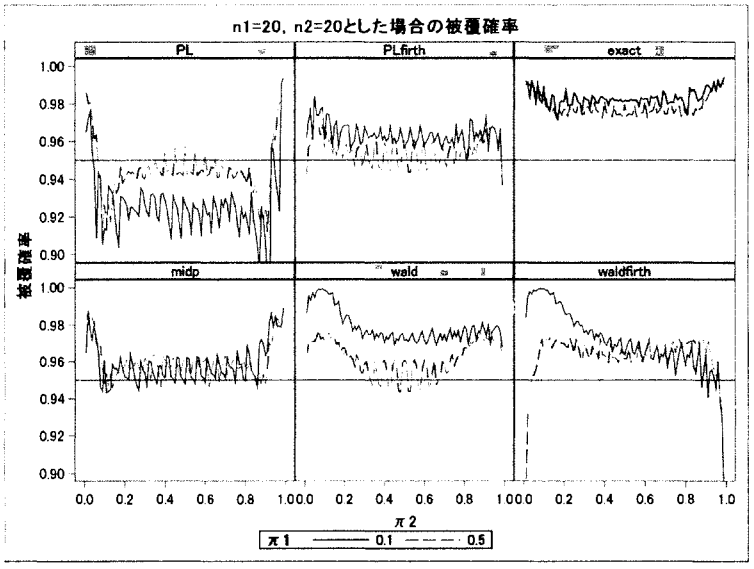


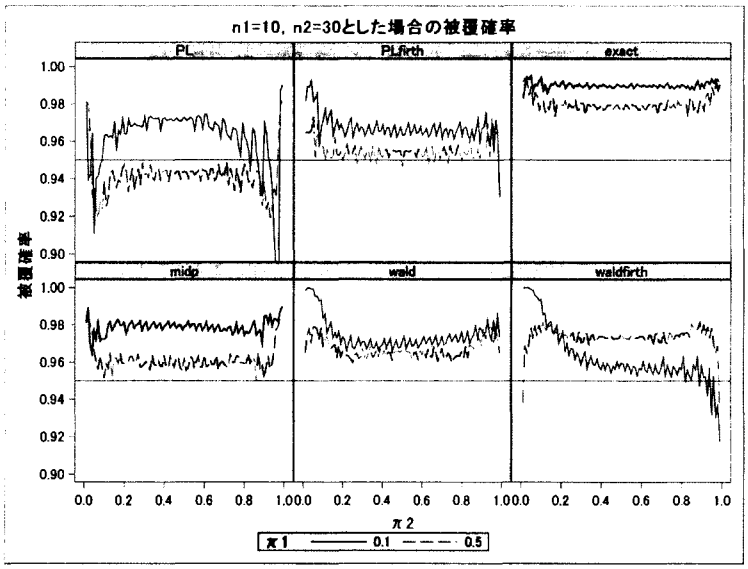
被覆確率を求める際の設定

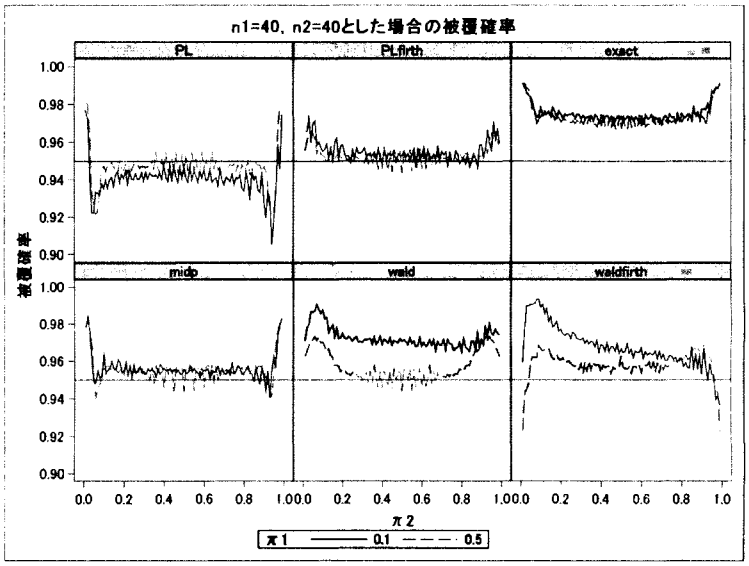
$$\left\{\begin{array}{cc} n_1 & n_2 \\ 20 & 20 \\ 10 & 30 \\ 40 & 40 \end{array}\right\} \times \left\{\begin{array}{c} \pi_1 \\ 0.1 \\ 0.5 \end{array}\right\} \times \left\{\begin{array}{c} \pi_2 \\ 0.01 \\ \vdots \\ 0.99 \end{array}\right\}$$

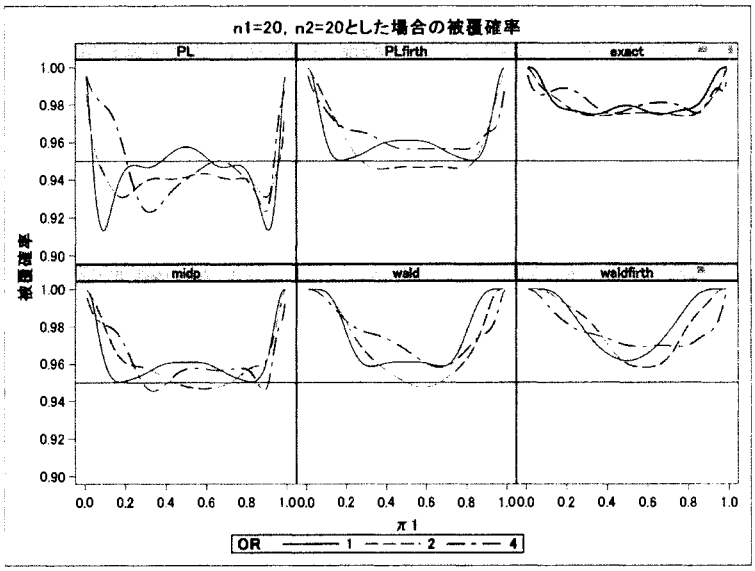
$$\left\{\begin{array}{cc} n_1 & n_2 \\ 20 & 20 \\ 10 & 30 \\ 40 & 40 \end{array}\right\} \times \left\{\begin{array}{c} \mathrm{OR} \\ 1 \\ 2 \\ 4 \end{array}\right\} \times \left\{\begin{array}{c} \pi_1 \\ 0.01 \\ \vdots \\ 0.99 \end{array}\right\}$$

信頼区間：両側95%水準

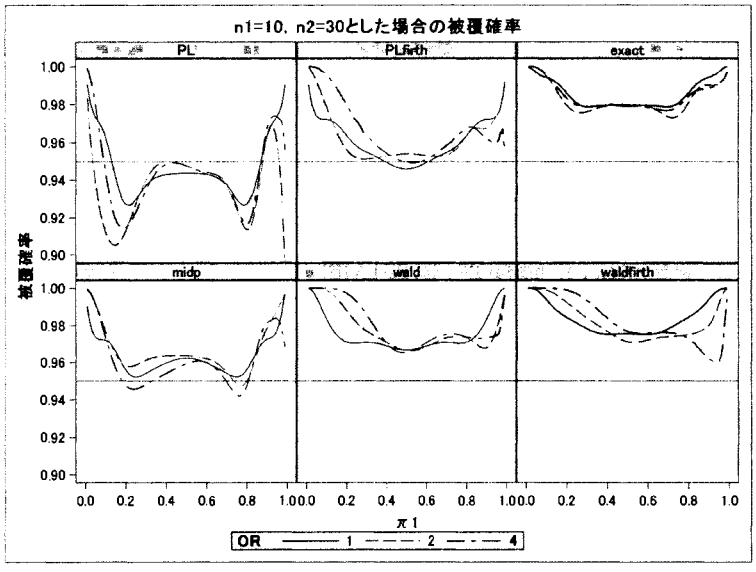


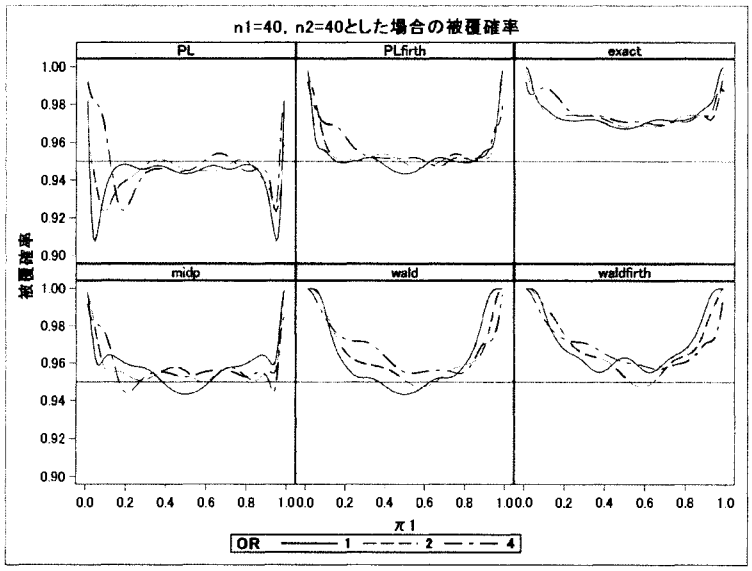

n1=10, n2=30とした場合の被覆確率
PL
PLfirth
exact
midp
wald
waldfirth
被覆確率
1.00
0.98
0.96
0.94
0.92
0.90
0.0
0.2
0.4
0.6
0.8
1.0
π 2
π 1
0.1
0.5


n1=40, n2=40とした場合の被覆確率
PL
PLfirth
exact
midp
wald
waldfirth
被覆確率
1.00
0.98
0.96
0.94
0.92
0.90
0.0
0.2
0.4
0.6
0.8
1.0
π 2
π 1
0.1
0.5



医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014





医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

まとめ

- 被覆確率が95%に近い信頼区間
  - PLFirth信頼区間
  - Midp信頼区間
- 保守的である信頼区間
  - Exact(正確検定)信頼区間

今後の課題

- 保守的でより名義水準に近くなるような信頼区間の提案
- 層を併合した場合の適切な共通オッズ比の信頼区間を評価する

参考文献

[1] A. Agresti. and Y. Min, *Unconditional small-sample confidence intervals for the odds ratio*, Biostatistics, 3, 3, 379-386, 2002.

[2] A. Agresti, *Categorical Data Analysis*, *Second Edition*: WILEY INTERSCIENCE, 2003.

[3] J. L. Fleiss, B. Levin, M. Cho Paik, *Statistical Methods for Rates and Proportions*, *Third Edition* : WILEY INTERSCIENCE, 2003.

[4] Newcombe, R. G. , *Confidence Intervals for Proportions and Related Measures of Effect Size*, Chapman & Hall/CRC, 2012.

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

参考文献

[5] OKUMURA. A. et al., *Hypocarbia in Preterm Infants With Periventricular Leukomalacia: The Relation Between Hypocarbia and Mechanical Ventilation*, Pediatrics, 107, 469–475, 2001.

[6] Woodward, *Epidemiology* : *Study Design and Data Analysis*, Chapman & Hall/CRC, 2004.

ネットワークメタアナリシスによる無作為化比較試験の統合

福井 伸行
株式会社データフォーシーズ

乙黒 俊也, 磯崎 充宏
日本たばこ産業株式会社

Combination of randomized controlled trials in Network Meta-Analysis

Nobuyuki Fukui
Data4C's K.K.

Toshiya Otoguro, Mitsuhiro Isozaki
Japan Tobacco Inc.

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

医療統計の分野で近年, ネットワークメタアナリシスという手法が注目を集めている。
本発表ではネットワークメタアナリシスのポイントを解説し, proc mcmcを用いた解析の実例を紹介する。

キーワード: メタアナリシス, ネットワークメタアナリシス, mcmc

Agenda

1. 背景
   - メタアナリシス再考
   - ネットワークメタアナリシスへの拡張
2. ネットワークメタアナリシスの概要
   - 直接比較と間接比較の統合
   - inconsistency
3. ネットワークメタアナリシスの統計モデル
   - consistency Model
   - inconsistency Model
4. 実践例
5. まとめ

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

Agenda

1. 背景
   - メタアナリシス再考
   - ネットワークメタアナリシスへの拡張

1. 背景
   ➢メタアナリシス再考

メタアナリシス

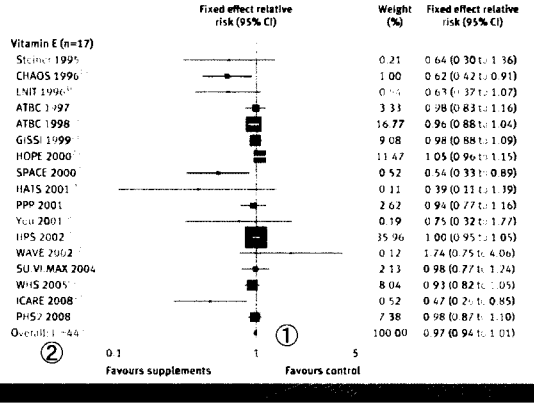
複数の無作為化比較試験の結果を集め,
1. 治療効果の重み付き平均を求める。(試験結果の統合を行う。)
2. 治療効果のバラつき(異質性)を検討する。

ための統計モデル。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

1. 背景
   ➢メタアナリシス再考

① 最下部のダイヤが推定値の統合結果を示している。
② 試験全体のバラつき(異質性)を統計量で定量的に評価することも可能。



Myung SK et.al, 2013

1. 背景
   ➢ネットワークメタアナリシスへの拡張

- メタアナリシスは高いエビデンスをもたらす優れた統計手法である。
- 通常のメタアナリシスは異なる試験の同じペアの比較結果を統合する。
- 異なる試験の異なるペアの比較結果を統合し、総合的に比較を行うことをモチベーションとしてネットワークメタアナリシスが提案された。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

Agenda

2. ネットワークメタアナリシスの概要
   - ➢ 直接比較と間接比較の統合
   - ➢ inconsistency

2. ネットワークメタアナリシスの概要
   - 直接比較と間接比較の統合

ネットワークメタアナリシス

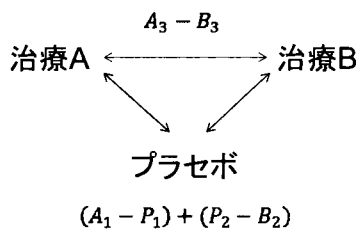
複数の無作為化比較試験を集め，直接比較と間接比較あるいは間接比較同士の統合を行うための統計モデル。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

2. ネットワークメタアナリシスの概要
   - 直接比較と間接比較の統合

（例）
プラセボ vs 治療A，プラセボ vs 治療B，治療A vs 治療Bの3つの無作為化比較試験の結果を統合する。

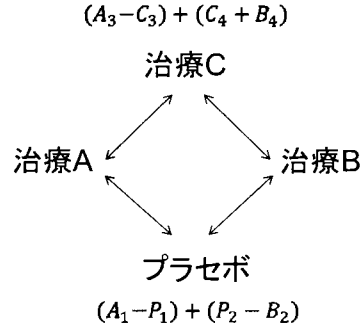
試験	プラセボ	治療A	治療B
1	P_1	A_1	
2	P_2		B_2
3		A_3	B_3



2. ネットワークメタアナリシスの概要
   - 直接比較と間接比較の統合

（例）
プラセボ vs 治療A, プラセボ vs 治療B, 治療A vs 治療C, 治療B vs 治療Cの4つの無作為化比較試験との結果を統合する。

	プラセボ	治療A	治療B	治療C
試験1	P_1	A_1		
試験2	P_2		B_2	
試験3		A_3		C_3
試験4			B_4	C_4

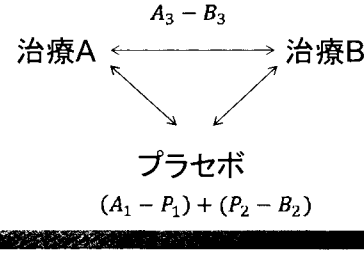


医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

2. ネットワークメタアナリシスの概要
   - inconsistency

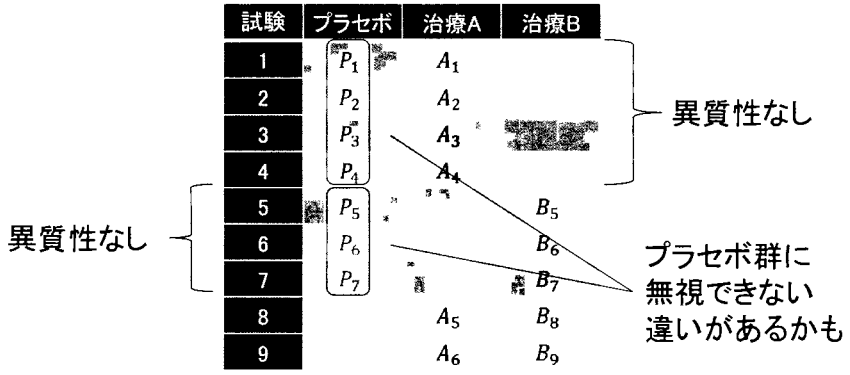
- 先の例では暗黙のうちに各試験の各治療群は同等と仮定。
- しかし、下の例で2つの試験のプラセボ群は同等であるとは限らない（試験の実施年代や除外基準の違い，背景因子のずれ）。
- AB間の比較の経路によって効果の差に不一致（inconsistency）が生じる可能性。

試験	プラセボ	治療A	治療B
1	P_1	A_1	
2	P_2		B_2
3		A_3	B_3



2. ネットワークメタアナリシスの概要
   - inconsistency

不一致性(inconsistency)は2群間の異質性(heterogeneity)とは独立に生じうる。



医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

2. ネットワークメタアナリシスの概要
   - inconsistency

ネットワークネタアナリシスの枠組みではinconsistencyの検出が可能。

次章で、
- 一致性を仮定したモデル
- 不一致性を検出するための基本的なモデル

を紹介する。

Agenda

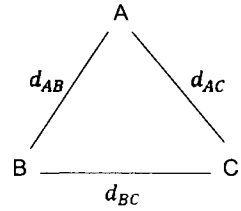
3. ネットワークメタアナリシスの統計モデル
   - consistency Model
   - inconsistency Model

3. ネットワークメタアナリシスの統計モデル
   - consistency model

- ネットワークメタアナリシスの基本となるモデル。
- 治療群間差に次のような一致性の仮定を置く。

(例) $d_{BC} = d_{AB} + d_{AC}$

試験	治療A	治療B	治療 C
1	$r_{1,A}/n_{1,A}$	$r_{1,B}/n_{1,B}$	
2		$r_{2,B}/n_{2,B}$	$r_{2,C}/n_{2,C}$
3	$r_{3,A}/n_{3,A}$		$r_{3,C}/n_{3,C}$
		⋮	



3. ネットワークメタアナリシスの統計モデル
   - consistency model

観測データは次のような確率分布に従うと仮定する。

試験	治療A	治療B	治療 C
1	$r_{1,A}/n_{1,A}$	$r_{1,B}/n_{1,B}$	[illegible]
2		$r_{2,B}/n_{2,B}$	$r_{2,C}/n_{2,C}$
3	$r_{3,A}/n_{3,A}$		$r_{3,C}/n_{3,C}$
		⋮	

$$
\begin{aligned}
&r_{1,A} \sim \text{Binomial}(p_{1,A}, n_{1,A}) \\
&r_{1,B} \sim \text{Binomial}(p_{1,B}, n_{1,B}) \\
&p_{1,A} = \text{logistic}(\mu_1) \\
&p_{1,B} = \text{logistic}(\mu_1 + \delta_{1,AB}) \\
&\delta_{1,AB} \sim \text{Normal}(d_{AB}, \sigma^2) \quad \text{変量効果}
\end{aligned}
$$

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

3. ネットワークメタアナリシスの統計モデル
   - consistency model

1試験に3つ以上の治療群があるときは注意が必要。

試験	治療A	治療B	治療 C
1	$r_{1,A}/n_{1,A}$	$r_{1,B}/n_{1,B}$	$r_{1,C}/n_{1,C}$
2		$r_{2,B}/n_{2,B}$	$r_{2,C}/n_{2,C}$
3	$r_{3,A}/n_{3,A}$		$r_{3,C}/n_{3,C}$
		⋮	

$$
\begin{aligned}
&r_{1,A} \sim \text{Binomial}(p_{1,A}, n_{1,A}) \\
&r_{1,B} \sim \text{Binomial}(p_{1,B}, n_{1,B}) \\
&r_{1,C} \sim \text{Binomial}(p_{1,C}, n_{1,C}) \\
&p_{1,A} = \text{logistic}(\mu_1) \\
&p_{1,B} = \text{logistic}(\mu_1 + \delta_{1,AB}) \\
&p_{1,C} = \text{logistic}(\mu_1 + \delta_{1,AC}) \\
&\begin{pmatrix} \delta_{1,AB} \\ \delta_{1,AC} \end{pmatrix} \sim N\left(\begin{pmatrix} d_{AB} \\ d_{AC} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma^2/2 \\ \sigma^2/2 & \sigma^2 \end{pmatrix} \right)
\end{aligned}
$$

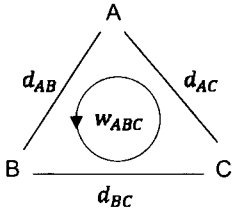
AB間の差とAC間の差に相関があるため，
多変量正規分布にしなければならない。

3. ネットワークメタアナリシスの統計モデル
   - inconsistency model

- 観測データが従う確率分布はconsistency modelと同じ。
- 治療群間差の一致性の仮定にパラメータ(inconsistency parameter)を追加する。
- このパラメータを調べることで，不一致性を評価する。

(例) $d_{BC} = d_{AB} + d_{AC} + w_{ABC}$

試験	治療A	治療B	治療 C
1	$r_{1,A}/n_{1,A}$	$r_{1,B}/n_{1,B}$	
2		$r_{2,B}/n_{2,B}$	$r_{2,C}/n_{2,C}$



医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

Agenda

4. 実践例

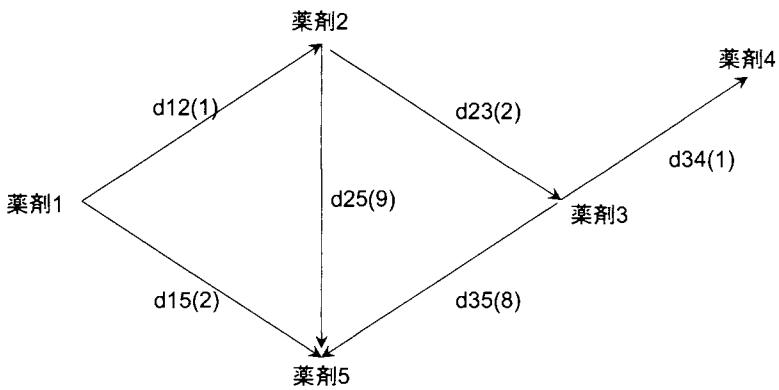
4. 実践例

ID	Study	薬剤1	薬剤2	薬剤3	薬剤4	薬剤5（プラセボ）	Combination	Design
1	Study 1		$r_{1,2}/n_{1,2}$	$r_{1,3}/n_{1,3}$		$r_{1,5}/n_{1,5}$		
2	Study 2		$r_{2,2}/n_{2,2}$			$r_{2,5}/n_{2,5}$		
3	Study 3		$r_{3,2}/n_{3,2}$			$r_{3,5}/n_{3,5}$		
4	Study 4			$r_{4,3}/n_{4,3}$		$r_{4,5}/n_{4,5}$		
5	Study 5		$r_{5,2}/n_{5,2}$			$r_{5,5}/n_{5,5}$		
6	Study 6			$r_{6,3}/n_{6,3}$		$r_{6,5}/n_{6,5}$		
7	Study 7		$r_{7,2}/n_{7,2}$	$r_{7,3}/n_{7,3}$		$r_{7,5}/n_{7,5}$		
8	Study 8							
9	Study 9							
10	Study 10			$r_{10,3}/n_{10,3}$		$r_{10,5}/n_{10,5}$		
11	Study 11			$r_{11,3}/n_{11,3}$		$r_{11,5}/n_{11,5}$		
12	Study 12		$r_{12,2}/n_{12,2}$			$r_{12,5}/n_{12,5}$		
13	Study 13			$r_{13,3}/n_{13,3}$		$r_{13,5}/n_{13,5}$		
14	Study 14			$r_{14,3}/n_{14,3}$		$r_{14,5}/n_{14,5}$		
15	Study 15			$r_{15,3}/n_{15,3}$	$r_{15,4}/n_{15,4}$			
16	Study 16	$r_{16,1}/n_{16,1}$	$r_{16,2}/n_{16,2}$			$r_{16,5}/n_{16,5}$		
17	Study 17	$r_{17,1}/n_{17,1}$				$r_{17,5}/n_{17,5}$		

当日の発表は数値データで行います。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

4. 実践例



4. 実践例

- proc mcmcを使って、最尤推定を実行する。
- 変量効果$\delta_{k,ij}$~Normal(d_{ij},σ^2)のパラメータd_{ij},σおよび inconsistency paramter wの分散に無情報事前分布を仮定する階層ベイズモデルとしてモデリング。

$$\begin{cases} d_{ij}\sim\text{Normal}(0,10^2) \\ \sigma\sim\text{gamma}(0.01,100) \\ \text{w}\sim\text{Normal}(0,\tau^2) \\ \tau\sim\text{gamma}(0.01,100) \end{cases}$$

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

4. 実践例

インプットデータセット

arm1-arm3: 薬剤の番号
response1-response3: 反応数
total1-total3: 例数
na: 各試験の治療群数

arm1	arm2	arm3	response1	response2	response3	total1	total2	total3	na
2	3	5	$r_{1,2}$	$r_{1,3}$	$r_{1,5}$	$n_{1,2}$	$n_{1,3}$	$n_{1,5}$	3
2	5	.	$r_{2,2}$	$r_{2,5}$	.	$n_{2,2}$	$n_{2,5}$	.	2
2	5	.	$r_{3,2}$	$r_{3,5}$	.	$n_{3,2}$	$n_{3,5}$	.	2
3	5	.	$r_{4,3}$	$r_{4,5}$	.	$n_{4,3}$	$n_{4,5}$	.	2
2	5	.	$r_{5,2}$	$r_{5,5}$	.	$n_{5,2}$	$n_{5,5}$	.	2
3	5	.	$r_{6,3}$	$r_{6,5}$	.	$n_{6,3}$	$n_{6,5}$	.	2
2	3	5	$r_{7,2}$	$r_{7,3}$	$r_{7,5}$	$n_{7,2}$	$n_{7,3}$	$n_{7,5}$	3
2	5	.	$r_{8,2}$	$r_{8,5}$	.	$n_{8,2}$	$n_{8,5}$	.	2
2	5	.	$r_{9,2}$	$r_{9,5}$	.	$n_{9,2}$	$n_{9,5}$	.	2
3	5	.	$r_{10,3}$	$r_{10,5}$	.	$n_{10,3}$	$n_{10,5}$	.	2
3	5	.	$r_{11,3}$	$r_{11,5}$	.	$n_{11,3}$	$n_{11,5}$	.	2
2	5	.	$r_{12,2}$	$r_{12,5}$	.	$n_{12,2}$	$n_{12,5}$	.	2
3	5	.	$r_{12,3}$	$r_{13,5}$	.	$n_{12,3}$	$n_{13,5}$	.	2
3	5	.	$r_{13,3}$	$r_{14,5}$	.	$n_{13,3}$	$n_{14,5}$	.	2
3	4	.	$r_{14,3}$	$r_{14,4}$	.	$n_{14,3}$	$n_{14,4}$	.	2
1	2	5	$r_{16,1}$	$r_{16,2}$	$r_{16,5}$	$n_{16,1}$	$n_{16,2}$	$n_{16,5}$	3
1	5	.	$r_{17,1}$	$r_{17,5}$	.	$n_{17,1}$	$n_{17,5}$	.	2

4. 実践例

proc mcmc オプション

```
proc mcmc data = ObservedData
          nmc = 10000
          nbi = 1000
          thin = 5
          missing = ac       /*欠測値を含むオブザベーションをモデルに含める*/
          dic                /*DICを出力する*/
          statistics(alpha = 0.05) = (summary interval);

～・～・～ ・～ ・～ ・～ ・～(後述)・～ ・～ ・～ ・～ ・～ ・～ ・～

run;
```

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

4. 実践例

事前分布の指定

```
prior d_1_5 ~ normal(0, sd = 10);
prior d_2_5 ~ normal(0, sd = 10);
prior d_3_5 ~ normal(0, sd = 10);
prior d_3_4 ~ n(0, sd = 10);
prior w_1_2_5 ~ normal(0, sd = tau);
prior w_2_3_5 ~ normal(0, sd = tau);
parms d_1_5 d_2_5 d_3_5 d_3_4
      w_1_2_5 w_2_3_5 0;

prior tau ~ gamma(0.01, scale = 100);
parms tau 1;

prior sigma ~ gamma(0.01, scale = 1000);
parms sigma 1;

array mu[17];
prior mu: ~ normal(0, sd = 10);
parms mu: 0;
```

$d_{15} \sim \text{Normal}(0, 10^2)$
$d_{25} \sim \text{Normal}(0, 10^2)$
$d_{35} \sim \text{Normal}(0, 10^2)$
$d_{34} \sim \text{Normal}(0, 10^2)$

$w_{125} \sim \text{Normal}(0, \tau^2)$
$w_{235} \sim \text{Normal}(0, \tau^2)$

$\tau \sim \text{gamma}(0.01, 100)$

$\sigma \sim \text{gamma}(0.01, 100)$

4. 実践例

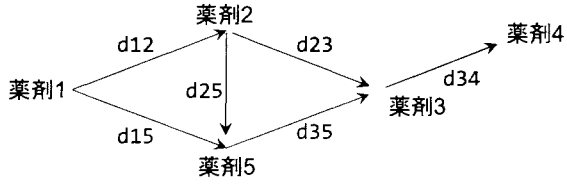
一致性の仮定, inconsistency parameterの導入

```
array d[5, 5] d11-d15 d21-d25 d31-d35 d41-d45 d51-d55;

do i = 1 to 5; do j = 1 to 5; d[i, j] = 0; end; end;

d12 = d_1_5 - d_2_5 + w_1_2_5;
d15 = d_1_5;
d23 = d_2_5 - d_3_5 + w_2_3_5;
d25 = d_2_5;
d34 = d_3_4;
d35 = d_3_5;
```

Inconsistency model の場合



4. 実践例

統計モデル①

```
md12 = d[arm1, arm2];
if na = 2 then md13 = 0;
else if na = 3 then md13 = d[arm1, arm3];

if na = 2 then do;
  S11 = sigma**2; S12 = 0;
  S21 = 0;        S22 = 1;
end;
else if na = 3 then do;
  S11 = sigma**2;     S12 = sigma**2 / 2;
  S21 = sigma**2 / 2; S22 = sigma**2;
end;

array vdelta[2] delta12 delta13;
array vmd[2] md12 md13;
array S[2, 2] S11 S12 S21 S22;

random vdelta ~ mvn(vmd, S) subject = _obs_;
```

多変量正規分布の
平均ベクトルの定義

多変量正規分布の
分散共分散行列の定義

$$\begin{pmatrix}\delta_{h,ij}\\ \delta_{h,ik}\end{pmatrix} \sim N\left(\begin{pmatrix}d_{ij}\\ d_{ik}\end{pmatrix}, \begin{pmatrix}\sigma^2 & \sigma^2/2\\ \sigma^2/2 & \sigma^2\end{pmatrix}\right)$$

4. 実践例

統計モデル②

```
p1 = logistic(mu[StudyNum]);
p2 = logistic(mu[StudyNum] + delta12);
p3 = logistic(mu[StudyNum] + delta13);

if na = 2 then
  lp = lpdfbin(response1, total1, p1)
       + lpdfbin(response2, total2, p2);
else if na = 3 then
  lp = lpdfbin(response1, total1, p1)
       + lpdfbin(response2, total2, p2)
       + lpdfbin(response3, total3, p3);

model general(lp);
```

$$p_{h,A} = \text{logistic}(\mu_h)$$
$$p_{h,B} = \text{logistic}(\mu_h + \delta_{h,AB})$$
$$p_{h,C} = \text{logistic}(\mu_h + \delta_{h,AC})$$

$$r_{h,A} \sim \text{Binomial}(p_{h,A}, n_{h,A})$$
$$r_{h,B} \sim \text{Binomial}(p_{h,B}, n_{h,B})$$
$$r_{h,C} \sim \text{Binomial}(p_{h,C}, n_{h,C})$$

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

4. 実践例

結果と考察

当日の発表で紹介します

Agenda

5. まとめ

5. まとめ

- ネットワークメタアナリシスは直接比較と間接比較あるいは間接比較同士を統合するための統計モデル。
- 直接比較と間接比較の結果を統合を考えるうえで，不一致性(inconsistency)を考慮に入れる必要がある。
- 一致性を仮定したうえで試験結果を統合するための consistency model と不一致性を評価するための inconsistency model による解析をSASのmcmcプロシージャを使って実施した。
- 一般的にネットワークメタアナリシスは結果の解釈が非常に難しい統計手法である。今後の研究の発展に期待したい。

参考文献

1. Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med 2002;21:2313-24.
2. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 2004;23:3105-24.
3. Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. *J Am Stat Assoc* 2006;101:477-59.
4. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med* 2010;29:932-44.
5. NICE DSU TECHNICAL SUPPORT DOCUMENT 2-5. NICE DECISION SUPPORT UNIT. http://www.nicedsu.org.uk/Evidence-Synthesis-TSD-series%282391675%29.htm
6. Gordon H. Guyatt et.al, The GRADE Working Group. GRADE guidelines: 8. Rating the quality of evidenced indirectness. *J Clin Epid* 2011;64:1303-1310.

FREQプロシジャによる割合の差の信頼区間

-V9.4における機能拡張と性能評価-

飯塚 政人[1]　魚住 龍史[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

Confidence intervals for the difference between proportions by FREQ procedure:
Enhancements in SAS 9.4 and performance evaluations

Masato Iizuka[1], Ryuji Uozumi[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

V9.3 ではFREQプロシジャのRISKDIFFによる割合の差の信頼区間として，8種類の方法による信頼区間が構成できた．そして今回新たにリリースされたV9.4においても，FREQプロシジャの機能が拡張され，RISKDIFFによる割合の差の信頼区間の構成方法として，3種類の方法（1)Agresti-Caffo信頼区間，2)Miettinen-Nurminen信頼区間，3)Mee信頼区間）が追加された．
本稿では，V9.4から新たに追加された3種類の構成方法を概説するとともに，被覆確率による性能評価を行い，V9.3までの信頼区間と比較する．さらに，ケース・スタディとして，優越性・同等性の臨床試験を想定した下，信頼区間幅やシミュレーションによる検出力について性能評価を行い，それぞれの場合において推奨すべき信頼区間の構成方法を報告する．

キーワード：FREQ，RISKDIFF，2項割合の差，信頼区間，被覆確率

1. はじめに

医薬研究の統計的評価をする際は，信頼区間の使用が推奨されている．医薬統計の教科書で，2群の割合の差の信頼区間の推定は，Wald 型の両側 95% 信頼区間を示していることが少なくない [5, 15]. 2 群の割合の差の信頼区間を考える上で，想定する2×2分割表を表1に示す．

表 1. 2×2 分割表

	有効	無効	計	有効割合	母数
薬剤群	n_{11}	n_{12}	$n_{1\cdot}$	$p_1 = \frac{n_{11}}{n_{1\cdot}}$	π_1
対照群	n_{21}	n_{22}	$n_{2\cdot}$	$p_2 = \frac{n_{21}}{n_{2\cdot}}$	π_2
計	$n_{\cdot 1}$	$n_{\cdot 2}$	n	$p = \frac{n_{\cdot 1}}{n}$	

このとき，2 群の有効割合の差の真値を$\Delta = \pi_1 - \pi_2$ ，その推定値を$\hat{\Delta} = p_1 - p_2$ と表す.

ここで，V9.4 で割合の差の信頼区間を出力させるための FREQ プロシジャの構文をプログラムに示し，FREQ プロシジャの TABLE ステートメントにおける RISKDIFF オプションにより構成できる割合の差の信頼区間を表 2 に示す．なお，正確な検定に基づく信頼区間は EXACT ステートメントの記述も必要となる.

プログラム ：FREQ プロシジャによる割合の差の信頼区間

```
proc freq data=data;
  tables group*response / riskdiff(cl = type);
  exact method;
run;
```

表 2. FREQ プロシジャにおける信頼区間の構成方法

信頼区間の構成方法	RISKDIFF (CL = *type*)	EXACT *method*	SAS Version
Wald 信頼区間 [2, 5]	WALD	–	○
Wald(連続修正) [2, 5]	WALD (CORRECT)	–	○
Hauck-Anderson 信頼区間 [7]	HA	–	○
Farrington－Manning 信頼区間 [4]	FM	–	○
Agresti-Caffo 信頼区間 [1]	AC | AGRESTICAFFO	–	☆
Mee 信頼区間 [8]	MN (CORRECT=MEE)	–	☆
Miettinen-Nurminen 信頼区間 [9]	MN | MN	–	☆
Newcombe スコア信頼区間 [11]	NEWCOMBE | SCORE | WILSON	–	○
Newcombe スコア（連続修正）信頼区間 [11]	NEWCOMBE | SCORE | WILSON (CORRECT)	–	○
正確な検定に基づく信頼区間 [13]	EXACT *	RISKDIFF	○
正確な検定に基づく信頼区間（FM スコア）[3]	EXACT *	RISKDIFF (FMSCORE)	○

○: V9.3 から利用可能，☆: V9.4 から新たに追加

V9.3 では 8 種類の方法による信頼区間が構成でき，飯塚，浜田 (2013) は，8 種類の信頼区間の性能評価を行い，Newcombe スコア信頼区間が最も良いと報告した [17]．そして今回新たにリリースされた V9.4 においても，FREQ プロシジャの機能が拡張され，新たに 3 種類の方法が追加された．

本稿では，V9.4 から新たに追加された 3 種類の構成方法を概説するとともに，被覆確率による性能評価を行い，V9.3 までの信頼区間と比較する．さらに，ケース・スタディとして，論文公表されている臨床試験（優越性試験，同等性試験）[6, 16] から得られた割合の数値を想定した下，信頼区間幅，モンテカルロシミュレーションによる検出力の性能評価を行い，それぞれの場合において推奨すべき信頼区間の構成方法を報告する．

第 2 節では，本研究で扱う 11 種類の信頼区間の数理を示す．第 3 節では，被覆確率の算出法，評価方法について述べる．第 4 節では，得られた結果から各信頼区間の特徴を明らかにし，第 5 節では，ケース・スタディとして，実データの割合を想定した下で評価を行う．そして第 6 節でまとめを示す．

2. 信頼区間の構成法の数理

2.1. Wald 信頼区間

Wald 信頼区間の構成は，Δの漸近正規性より以下の信頼区間で表される構成法．

$$\hat{\Delta} \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_{1\cdot}}+\frac{p_2(1-p_2)}{n_{2\cdot}}} \tag{1}$$

Wald 信頼区間の特徴は，Δのとりえる範囲である[-1,1]を超えて上限もしくは下限が形成されることがある．また，$n_{11}=n_{21}=0$，$n_{12}=n_{22}=0$のとき信頼区間が（0，0）となる．

2.2. Wald(連続修正)信頼区間

Wald 信頼区間に連続性の修正を加えた構成法．

$$\hat{\Delta} \pm \left(\frac{\frac{1}{n_{1\cdot}}+\frac{1}{n_{2\cdot}}}{2}+z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_{1\cdot}}+\frac{p_2(1-p_2)}{n_{2\cdot}}}\right) \tag{2}$$

Wald 信頼区間の特徴と同様に，Δのとりえる範囲である[-1,1]を超えて上限もしくは下限が形成されることがある．

2.3. Hauck-Anderson 信頼区間

Hauck-Anderson 信頼区間は，Wald 信頼区間より分散を大きくし，連続性の修正を加えた構成法．

$$\hat{\Delta} \pm \left(\frac{1}{2min(n_{1\cdot},n_{2\cdot})}+z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_{1\cdot}-1}+\frac{p_2(1-p_2)}{n_{2\cdot}-1}}\right) \tag{3}$$

Wald 信頼区間の特徴と同様に，Δのとりえる範囲である[-1,1]を超えて上限もしくは下限が形成されることがある．

2.4. Farrington-Manning 信頼区間

Farrigton-Manning 信頼区間は，帰無仮説$H_0: \pi_1 = \pi_2$の下で分散を考えている構成法.

$$\hat{\Delta} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n_{1\cdot}} + \frac{p(1-p)}{n_{2\cdot}}} \tag{4}$$

Wald 信頼区間の特徴と同様に，Δのとりえる範囲である[-1,1]を超えて上限もしくは下限が形成されることがある．また，$n_{11} = n_{21} = 0$，$n_{12} = n_{22} = 0$のとき信頼区間が（0，0）となる.

2.5. Agresti-Caffo 信頼区間

Agresti-Caffo 信頼区間の構成は，2 × 2 分割表の各セルに 1 度数足して，Wald 信頼区間を導いた構成法,

$$\hat{\Delta} \pm z_{\alpha/2}\sqrt{\frac{p_1^*(1-p_1^*)}{n_{1\cdot}+2} + \frac{p_2^*(1-p_2^*)}{n_{2\cdot}+2}} \tag{5}$$

$$p_1^* = \frac{n_{11}+1}{n_{1\cdot}+2}, p_2^* = \frac{n_{21}+1}{n_{2\cdot}+2} \tag{6}$$

Agresti-Caffo 信頼区間の特徴は,Δのとりえる範囲である[-1,1]を超えて上限もしくは下限が形成されることがある.

2.6. Mee 信頼区間

Mee 信頼区間の構成は，スコア型の信頼区間として以下のように考える.

$$\left|\hat{\Delta} - \Delta\right| \le z_{\alpha/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_{1\cdot}} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_{2\cdot}}} \tag{7}$$

$\tilde{p}_1$と$\tilde{p}_2$は制限付き最尤推定値であり，3 次方程式を解くことによって得られる.

$$\sum_{k=0}^{3} L_k p_1^k = 0 \tag{8}$$

ここで，$\tilde{p}_2 = \tilde{p}_1 + \Delta$と表せ，$L_3 = n$，$L_2 = (n_{2\cdot} + 2n_{1\cdot})\Delta - n - n_{11} - n_{21}$，$L_1 = (n_{1\cdot}\Delta - n - 2n_{11})\Delta + n_{11} + n_{21}$，$L_0 = n_{11}\Delta(1-\Delta)$である.

2.7. Miettinen-Nurminen 信頼区間

Miettinen-Nurminen 信頼区間の構成は，Mee の信頼区間に例数を加味し分散を大きくした構成法.

$$\hat{\Delta} \pm z_{\alpha/2}\sqrt{\left(\frac{n}{n-1}\right)\left(\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_{1\cdot}} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_{2\cdot}}\right)} \tag{9}$$

2.8. Newcombe スコア信頼区間

Newcombe スコア信頼区間は，単群の割合の Wilson スコア信頼区間 [13] に基づいた構成法.

$$(\text{上限}) = \hat{\Delta} + \sqrt{(U_1 - p_1)^2 + (p_2 - L_2)^2} \tag{10}$$

$$(\text{下限}) = \hat{\Delta} + \sqrt{(p_1 - L_1)^2 + (U_2 - p_2)^2} \tag{11}$$

ここでL_1とU_1は，単群の割合の Wilson スコア信頼区間の上限と下限を表し，以下のように求める．

$$|\pi_1 - p_1| = z_{\alpha/2}\sqrt{\frac{\pi_1(1-\pi_1)}{n_{1\cdot}}} \tag{12}$$

同様にL_2とU_2は，もう片方の群の割合の Wilson スコア信頼区間の上限と下限を表す．

$$|\pi_2 - p_2| = z_{\alpha/2}\sqrt{\frac{\pi_2(1-\pi_2)}{n_{2\cdot}}} \tag{13}$$

2.9. Newcombe スコア（連続修正）信頼区間

Newcombe スコア（連続修正）信頼区間は，単群の割合の Wilson スコア（連続修正）信頼区間 [13] に基づいた構成法．Newcombe スコア信頼区間の上限(10)，下限(11)の構成法は同様だが，L_1とU_1は，単群の割合の Wilson スコア（連続修正）信頼区間の上限と下限を表し，以下のように求める．

$$|\pi_1 - p_1| - \frac{1}{2n_{1\cdot}} = z_{\alpha/2}\sqrt{\frac{\pi_1(1-\pi_1)}{n_{1\cdot}}} \tag{14}$$

同様にL_2とU_2は，もう片方の群の割合の Wilson スコア（連続修正）信頼区間の上限と下限を表す．

$$|\pi_2 - p_2| - \frac{1}{2n_{2\cdot}} = z_{\alpha/2}\sqrt{\frac{\pi_2(1-\pi_2)}{n_{2\cdot}}} \tag{15}$$

2.10. 正確な検定に基づく信頼区間

正確な検定に基づく信頼区間の構成は，以下のように考える．

$$\text{（上限）} = \inf\left(\Delta: P_L(\Delta) > \frac{\alpha}{2}\right) \tag{16}$$

$$\text{（下限）} = \sup\left(\Delta: P_U(\Delta) > \frac{\alpha}{2}\right) \tag{17}$$

P_LとP_Uは，

$$P_U(\Delta) = \sup_{\pi_2}\left(\sum_{A,T\le t_0} f(n_{11}, n_{21}|\Delta, \pi_2)\right) \tag{18}$$

$$P_L(\Delta) = \sup_{\pi_2}\left(\sum_{A,T\ge t_0} f(n_{11}, n_{21}|\Delta, \pi_2)\right) \tag{19}$$

で表せられ，A: 2 × 2 分割表$(n_{1\cdot}, n_{2\cdot})$, T: 2 × 2 分割表の任意の値の検定統計量, t_0: 観測された値の検定統計量を表している．ここで用いられている検定統計量は，

$$T: p_1 - p_2 \tag{20}$$

となる．また$f(n_{11}, n_{21}|\Delta, \pi_2)$は，2 × 2分割表の同時確率を表しており，以下のように表される．

$$f(n_{11}, n_{21}|\Delta, \pi_2) = \binom{n_{1\cdot}}{n_{11}}(\Delta+\pi_2)^{n_{11}}(1-\Delta-\pi_2)^{n_{1\cdot}-n_{11}}\binom{n_{2\cdot}}{n_{21}}\pi_2^{n_{21}}(1-\pi_2)^{n_{2\cdot}-n_{21}} \tag{21}$$

正確な検定に基づく信頼区間の特徴は，全てのπ_2について名義有意水準以下となる．

2.11. 正確な検定（FM スコア）に基づく信頼区間

正確な検定（FM スコア）に基づく信頼区間は，（15）の検定統計量をスコア型（7）に変えた構成法．

$$T:\frac{p_1-p_2-\Delta}{\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_{1\cdot}}+\frac{\tilde{p}_2(1-\tilde{p}_2)}{n_{2\cdot}}}} \tag{22}$$

3. 評価基準

3.1. 被覆確率

信頼区間の性能を評価する際の主な評価指標として被覆確率を用いる．信頼区間の被覆確率（coverage probability）とは，信頼区間が対象の真値を含む確率のことである．被覆確率$\mathrm{CP}(\pi_1,\pi_2)$は以下の式で表される．

$$\begin{aligned}\mathrm{CP}(\pi_1,\pi_2) &= \sum_{n_{11}=0}^{n_{1\cdot}}\sum_{n_{21}=0}^{n_{2\cdot}} I(L\le\Delta\le U)\binom{n_{1\cdot}}{n_{11}}\pi_1{}^{n_{11}}(1-\pi_1)^{n_{1\cdot}-n_{11}}\binom{n_{2\cdot}}{n_{21}}\pi_2^{n_{21}}(1-\pi_2)^{n_{2\cdot}-n_{21}} \\ &= \sum_{n_{11}=0}^{n_{1\cdot}}\sum_{n_{21}=0}^{n_{2\cdot}} I(L\le\Delta\le U)\,f(n_{11},n_{21}|\pi_1,\pi_2)\end{aligned} \tag{23}$$

$I(L\le\Delta\le U)$は，指示関数であり，割合の差の真値Δが上限Uと下限Lの間に入っていれば 1 となり，その他は 0 となる関数である．信頼区間の幅を広げれば被覆確率は 1 に近づき，逆に狭めることで 0 に近づくが，被覆確率は信頼区間の名義水準に近くなることが望ましい．両側水準αの信頼区間を求めたときに，被覆確率が$(1-\alpha)$よりも小さい場合には，真値を含まない確率がαより大きく，英語でいうリベラルの対訳として革新的と表現する．一方，被覆確率が$(1-\alpha)$より大きい場合には，信頼区間の幅が必要以上に広いため，保守的と表現する．本稿では，被覆確率を用いて，2 種類の基準で信頼区間の構成方法の性能を評価した．1 つ目は，被覆確率が$(1-\alpha)$に近いことである．2 つ目は，実際は有意でないのに有意であると判定してしまうαエラーを常に一定以下に保持するように，被覆確率が常に$(1-\alpha)$よりも大きいことである．

3.2. 信頼区間幅

被覆確率が各信頼区間の構成法でほぼ等しい場合，信頼区間の幅が小さい方が推定精度の点において，好まれる．信頼区間幅は，第 5 節における評価に用いる．

3.3. 検出力（Power）

対立仮説が真の場合，帰無仮説を正しく棄却する確率を表す．第一種の過誤確率が名義水準に保たれている場合，検出力は高いほど良い．そして通常の第Ⅲ相臨床試験において，検出力は 80%以上に設定される．信頼区間幅同様，検出力も第 5 節における評価に用いる．

$$\mathrm{Power}=\frac{\text{有意になった回数}}{\text{シミュレーション回数}} \tag{24}$$

4. 割合の差の信頼区間の構成法の比較

図1, 図2は, 各群のサンプルサイズの総数nを40とし,$n_1 = n_2 = 20$とした場合(図1)と$n_1 = 10, n_1 = 30$とした場合（図2）で，$\pi_1 = 0.1,0.5$，$\pi_2 = 0.01,0.02,\ldots,099$のように条件を変え，95%信頼区間を形成したときの被覆確率を示した図である．また表3は，$\pi_1 = 0.01,0.02,\ldots,099$，$\pi_2 = 0.01,0.02,\ldots,099$の下，サンプルサイズが等しい場合の$n_1 = n_2 = 20$，$n_1 = n_2 = 50$，そして異なる場合の$n_1 = 10, n_2 = 30$の被覆確率の平均値と最小値を示した図である．

FREQプロシジャのデフォルトで表示されるWald信頼区間は，被覆確率が非常に小さくなっていることがわかる．また連続修正を加味したWald（連続修正）信頼区間も，真値π_1が0や1に近くなると被覆確率が名義水準から大きく外れてしまう場合がある．この傾向は，総数nが同じで各群のサンプルサイズが異なった場合でも同様であり，より顕著となる．

表3から，被覆確率が名義水準である95%に近い信頼区間としては，スコア型の信頼区間である，Mee信頼区間，Miettinen-Nurminen信頼区間であることがわかる．またNewcombeスコア信頼区間も被覆確率が名義水準である95%に近い．Newcombeスコア信頼区間は，各群のサンプルサイズが異なった場合でも被覆確率の値にあまり影響をうけることなく，各群のサンプルサイズが等しい場合とあまり変わらない．ただし，π_1, π_2の値が0や1に近づくと名義水準を外れてしまうことがある．

保守的である信頼区間の構成法は，表3の被覆確率の最小値（Min）より，正確な検定に基づく信頼区間，正確な検定（FMスコア）に基づく信頼区間，Newcombeスコア（連続修正）信頼区間の3つである．各群のサンプルサイズが等しい場合は，正確な検定（FMスコア）に基づく信頼区間が保守的かつ名義水準に近い値となっている．各群のサンプルサイズがアンバランスな場合は，単群試験などで真値が0.5付近と予想される場合には，正確な検定に基づく信頼区間が名義水準に近い値となっており推奨される．しかし真値が0や1に近い値となると，極端に名義水準を外れてしまう傾向があるので，群の真値に関する情報が乏しい場合は正確な検定（FMスコア）に基づく信頼区間の使用が推奨される．

表3．サンプルサイズ別被覆確率の平均値（Mean）と最小値（Min）

	$n_1. = n_2. = 20$		$n_1. = n_2. = 50$		$n_1. = 10$ $n_2. = 30$	
	Mean	Min	Mean	Min	Mean	Min
Agresti-Caffo	**0.955**	0.936	**0.952**	0.942	**0.958**	0.917
Farrington-Manning	0.962	0.912	0.964	0.932	**0.949**	0.785
Hauck-Anderson	**0.957**	0.884	**0.958**	0.918	**0.949**	0.740
Miettinen-Nurminen	**0.951**	0.932	**0.950**	0.931	**0.956**	0.926
Mee	**0.948**	0.917	**0.949**	0.931	**0.953**	0.926
Newcombeスコア	**0.951**	0.920	**0.951**	0.934	**0.952**	0.928
Newcombeスコア(修正)	0.974	**0.951**	0.967	**0.955**	0.977	**0.957**
Wald	0.930	0.805	**0.942**	0.889	0.899	0.659
Wald(修正)	0.969	0.900	0.967	0.944	**0.955**	0.740
正確	0.978	**0.952**	0.974	**0.955**	0.978	**0.955**
正確(FMスコア)	0.963	**0.951**	**0.959**	**0.951**	0.975	**0.957**

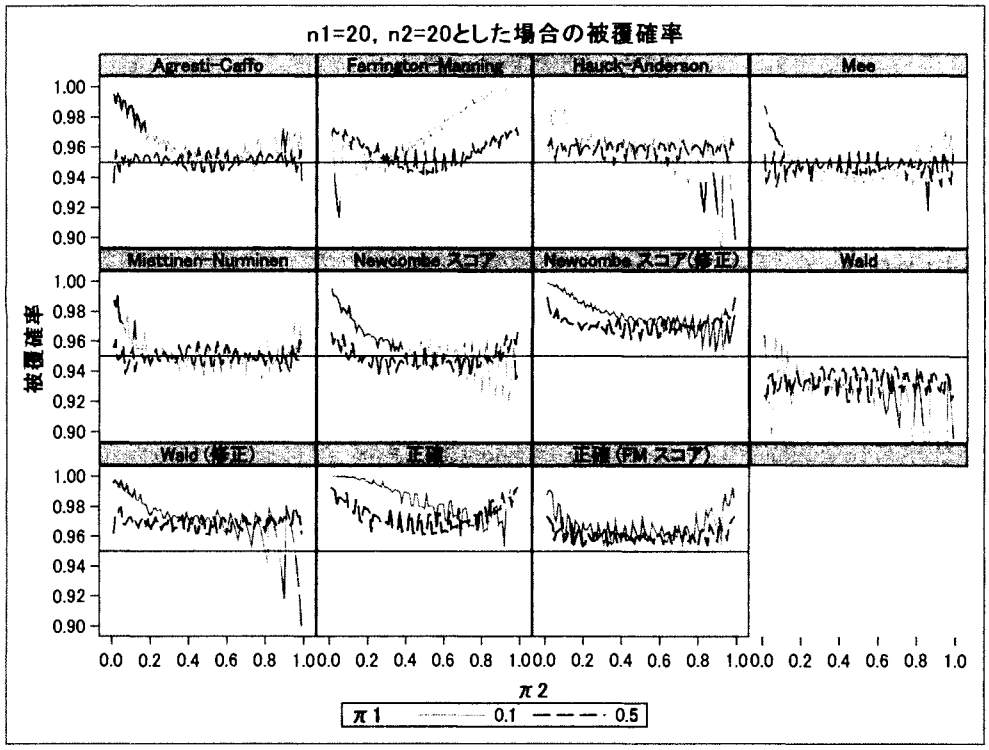


図 1. 各群のサンプルサイズを 20 とした被覆確率

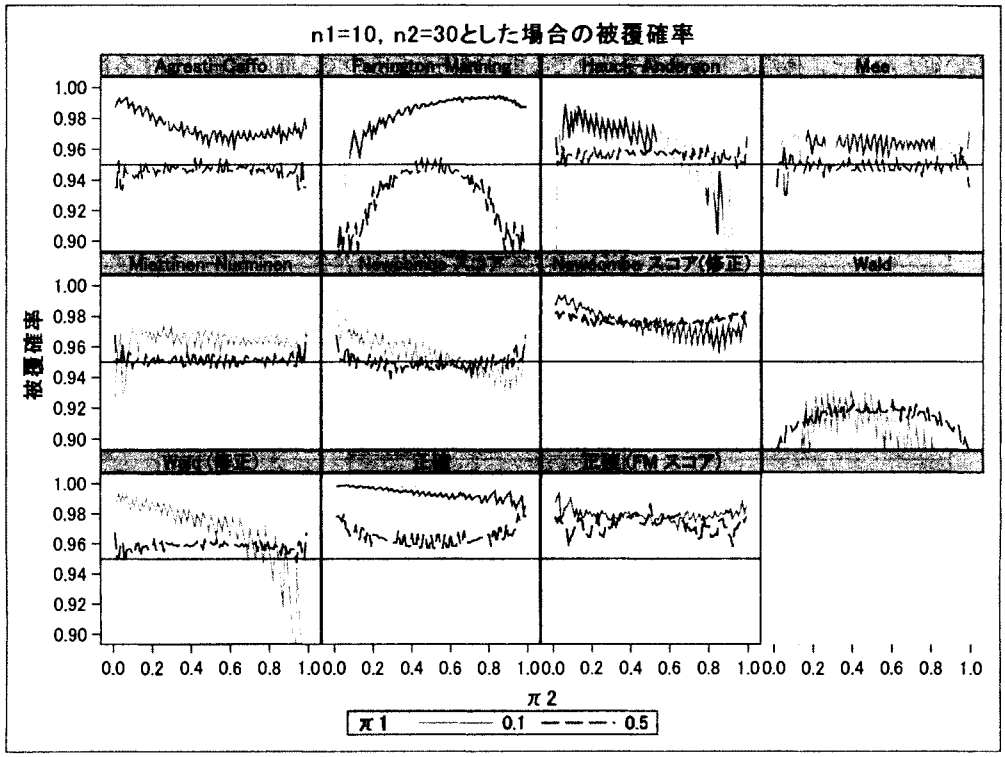


図 2. 各群のサンプルサイズを 10, 30 とした被覆確率

5. ケース・スタディ

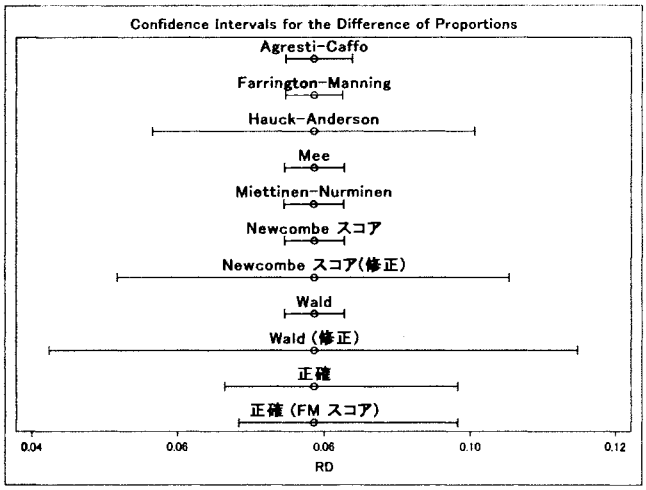
臨床試験から得られた 2 値データに対して統計解析を行う場合，非劣性試験や同等性試験においては，信頼区間に基づく意思決定を行う．そこで本節では，i) 前節で検討した例数と同程度の優越性試験から得られたデータ，ii) 同等性試験から得られたデータ，に基づき構成した 11 種類の信頼区間に対する評価を行う．

5.1. 優越性試験

表 4 は，関節リウマチ患者を対象とした第 II 相試験から得られた，ある有効性評価項目の結果の一部を示す [6]. 図 3 は，表 4 のデータから構成した 11 種類の 95%信頼区間を図示したもの及び，各信頼区間の上限，下限，区間幅を具体的な数値として示したものである．示した図表より，構成法によって区間の広がり方が異なっていることがわかる．特に連続修正を加味している構成法の区間幅が広く，最も区間幅が狭い Farrington-Manning 信頼区間と比べると区間幅の違いが顕著である．また，正確な検定に基づく構成法も区間幅が広い．Miettinen-Nurminen 信頼区間，Newcombe スコア信頼区間，Wald 信頼区間，Mee 信頼区間の幅は狭いことがわかる．

表 4. 優越性試験から得られた結果

	有効	無効	計	有効割合	差
試験群	3	25	28	0.107	0.078
対照群	1	34	35	0.029	



信頼区間	信頼下限	信頼上限	区間幅
Agresti-Caffo	0.0747	0.0838	0.0091
Farrington-Manning	0.0747	0.0824	0.0078
Hauck-Anderson	0.0566	0.1006	0.0440
Miettinen-Nurminen	0.0745	0.0827	0.0082
Newcombe スコア	0.0745	0.0827	0.0081
Wald	0.0745	0.0826	0.0081
Newcombe スコア(修正)	0.0517	0.1053	0.0536
Wald (修正)	0.0424	0.1148	0.0724
Mee	0.0745	0.0827	0.0081
正確	0.0664	0.0983	0.0318
正確 (FM スコア)	0.0684	0.0983	0.0299

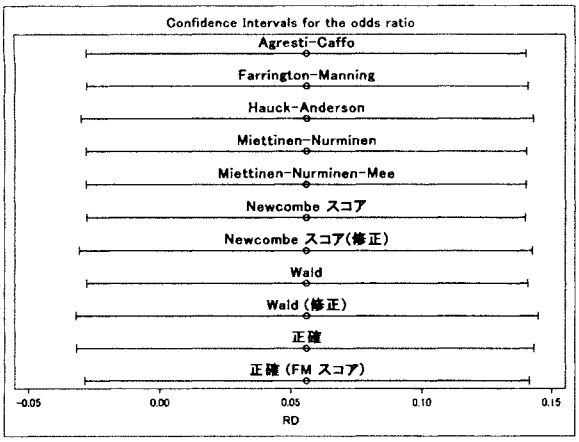
図 3. 95%信頼区間及び区間幅

5.2. 同等性試験

表 5 は，関節リウマチ患者を対象とした第 III 相試験から得られた，ある有効性評価項目の結果の一部を示す [16]. 同等性マージンを [−15%, 15%] として，有効割合の差の 95%信頼区間が同等性マージン内に含まれれば，同等と意思決定を行う．

表 5. 同等性試験から得られた結果

	有効	無効	計	有効割合	差
試験群	98	150	248	0.395	0.056
対照群	85	166	251	0.339	



信頼区間	信頼下限	信頼上限	区間幅
Agresti-Caffo	-0.028	0.140	0.168
Farrington-Manning	-0.028	0.141	0.169
Hauck-Anderson	-0.030	0.143	0.173
Mee	-0.028	0.141	0.169
Miettinen-Nurminen	-0.028	0.140	0.168
Newcombe スコア	-0.028	0.140	0.168
Newcombe スコア(修正)	-0.031	0.143	0.173
Wald	-0.028	0.141	0.169
Wald (修正)	-0.032	0.145	0.177
正確	-0.032	0.143	0.175
正確 (FM スコア)	-0.029	0.141	0.170

図 4. 95%信頼区間及び区間幅

図 4 は，表 5 のデータから構成した 11 種類の 95%信頼区間を図示したもの及び，信頼区間の上限，下限，区間幅を具体的な数値として示したものである．示した図表より，構成法によって区間の広がり方と区間幅の位置が若干異なっていることがわかる．特に連続修正を加味している構成法と正確な検定に基づく構成法の幅が比較的に広くなっている．なお，本ケース・スタディでは例数が多くかつ 40%近い有効割合のため，その他では区間幅に大きな違いはみられない．

次に，表 6 に様々な条件の下で求めた 10000 回のシミュレーションによる検出力を示す．P1，P2 は各群における有効割合，n1，n2 は各群の例数を表す．同等性試験なので，P1 と P2 が近い値を想定した下でシミュレーションを行った．具体的に述べると，対照群の有効割合 P2 は試験群の有効割合 P1 より大きくなることはないという想定の下（P1≥P2），P1=0.1，0.3，0.5 とし，P2 は P1 − (0 or 0.05 or 0.1) を想定した．また，Agresti-Caffo 信頼区間，Farrington－Manning 信頼区間 ，Hauck-Anderson 信頼区間 ，Miettinen-Nurminen 信頼区間，Newcombe スコア，Newcombe スコア（連続修正），Wald(連続修正)信頼区間をそれぞれ，AC，FM，HA，MN，NS，NSC，WALDC と表記する．なお，正確な検定に基づく信頼区間については，コンピュータの性能より求められないため，記述していない．

表 6 より，P1，P2 の値が小さく，例数が少ない場合は，Wald 信頼区間の検出力が上がることがわかる．他の条件の場合では，Agresti-Caffo 信頼区間，Newcombe スコア信頼区間の検出力が他の信頼区間に比べて高いことがわかる．また，サンプルサイズの議論となるが，例数が太字になっている箇所は，検出力が 80%超えているところである．割合の分散が大きくなる有効割合の真値 0.5 付近では，検出力 80%を超えるためには，例数を多く必要とすることがわかる．

表 6. 各条件の下で求めた信頼区間別の検出力

P1	P2	n1	n2	AC	FM	HA	MN	MEE	NS	NSC	WALD	WALDC
0.1	0.05	100	100	0.75	0.75	0.72	0.69	0.69	0.69	0.64	**0.77**	0.69
0.1	0.05	**248**	**251**	0.98	0.98	0.98	0.98	0.98	0.98	0.97	0.98	0.97
0.1	0.05	**300**	**300**	1.00	1.00	0.99	0.99	0.99	0.99	0.99	1.00	0.99
0.1	0.05	**500**	**500**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.1	0.1	**100**	**100**	**0.88**	0.87	0.85	0.84	0.84	0.85	0.80	**0.88**	0.81
0.1	0.1	**248**	**251**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.1	0.1	**300**	**300**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.1	0.1	**500**	**500**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.3	0.2	100	100	**0.12**	0.11	0.09	0.11	0.11	**0.12**	0.09	0.11	0.08
0.3	0.2	248	251	0.25	0.25	0.23	0.25	0.25	0.25	0.23	0.25	0.22
0.3	0.2	300	300	**0.30**	0.29	0.27	0.29	0.29	**0.30**	0.27	0.29	0.26
0.3	0.2	500	500	**0.46**	0.45	0.44	0.45	0.45	**0.46**	0.44	0.45	0.43
0.3	0.25	100	100	**0.25**	0.24	0.18	0.24	0.24	**0.25**	0.19	0.24	0.16
0.3	0.25	248	251	0.71	0.71	0.69	0.71	0.71	0.71	0.69	0.71	0.67
0.3	0.25	300	300	0.79	0.78	0.77	0.79	0.79	0.79	0.77	0.79	0.76
0.3	0.25	**500**	**500**	0.95	0.94	0.94	0.94	0.95	0.95	0.94	0.94	0.94
0.3	0.3	100	100	**0.31**	0.29	0.21	0.30	0.30	**0.31**	0.23	0.29	0.17
0.3	0.3	**248**	**251**	**0.92**	0.91	0.90	0.91	0.91	**0.92**	0.90	0.91	0.89
0.3	0.3	**300**	**300**	0.96	0.96	0.95	0.96	0.96	0.96	0.95	0.96	0.95
0.3	0.3	**500**	**500**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.5	0.4	100	100	0.06	0.06	0.02	0.06	0.06	0.06	0.02	0.06	0.02
0.5	0.4	248	251	0.20	0.20	0.18	0.20	0.20	0.20	0.19	0.20	0.18
0.5	0.4	300	300	0.24	0.24	0.21	0.24	0.24	0.24	0.22	0.24	0.21
0.5	0.4	500	500	0.36	0.36	0.34	0.36	0.36	0.36	0.34	0.36	0.34
0.5	0.45	100	100	0.13	0.13	0.04	0.13	0.13	0.13	0.04	0.13	0.04
0.5	0.45	248	251	0.61	0.61	0.58	0.61	0.61	0.61	0.58	0.61	0.58
0.5	0.45	300	300	0.70	0.69	0.67	0.70	0.70	0.70	0.67	0.70	0.67
0.5	0.45	**500**	**500**	0.89	0.89	0.88	0.89	0.89	0.89	0.88	0.89	0.88
0.5	0.5	100	100	0.17	0.17	0.06	0.17	0.17	0.17	0.06	0.17	0.05
0.5	0.5	**248**	**251**	0.84	0.84	0.82	0.84	0.84	0.84	0.82	0.84	0.82
0.5	0.5	**300**	**300**	0.92	0.91	0.90	0.92	0.92	0.92	0.90	0.92	0.90
0.5	0.5	**500**	**500**	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99

まとめ

本稿では，SAS の V9.4 の FREQ プロシジャから新たに追加された 3 種類の構成方法を含めた 11 種類の信頼区間の構成方法の特徴について，被覆確率を評価指標として比較した．その結果，被覆確率が名義水準に平均的に近かったのは， Miettinen-Nurminen 信頼区間と Newcombe スコア信頼区間であった．また，ケース・スタディの結果から，検出力が比較的高い信頼区間は，Newcombe スコアであったため，Newcombe スコア信頼区間が推奨される信頼区間であることが確認された．保守的な信頼区間は，正確な検定に基づく信頼区間，正確な検定（FM スコア）に基づく信頼区間，Newcombe スコア（連続修正）信頼区間であった．保守的な信頼区間の中でも被覆確率が名義水準に近かった信頼区間は，正確な検定（FM スコア）に基づく信頼区間であったため，保守的な信頼区間の中で推奨される信頼区間は正確な検定（FM スコア）に基づく信頼区間で

あった．しかし，各群において例数が異なる場合で，単群等の試験などで真値が 0.5 付近であると予想できる場合には，正確な検定に基づく信頼区間の使用も検討される．

参考文献

[1] Agresti, A. and Caffo,. B., Simple and effective confidence intercals for proportions and differences of proportions result from adding two successes and two failures, *The American Statistician*, 54, 280-288, 2000.

[2] Agresti, A. *Categorical Data Analysis*, *Second Edition*: John Wiley & Sons, 2003.

[3] Chan, I. S. F., Zhang, Z., Test-based exact confidence intervals for the difference of two binomial proportions, *Biometrics*, 55, 1202-1209, 1999.

[4] Farrington, C. P. and Manning, G., Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk, *Statistics in Medicine*, 9, 1447–1454, 1990.

[5] Fleiss, J. L., Levin, B., Paik, M. C., *Statistical Methods for Rates and Proportions*, *Third Edition* : John Wiley & Sons, 2003.

[6] Genovese. M.C. et al., Tabalumab, an anti-BAFF monoclonal antibody, in patients with active rheumatoid arthritis with an inadequate response to TNF inhibitors. *Annals of the Rheumatic Diseases*, 72, 1461–1468, 2013.

[7] Hauck, W. W. and Anderson, S., A comparison of large-sample confidence interval methods for the difference of two binomial probabilities, *The American Statistician*, 40, 318–322, 1986.

[8] Mee, R. W., Confidence bounds for the difference between two probabilities, *Biometrics*, 40, 1175-1176, 1984.

[9] Miettinen, O. S. and Nurminen, M., Comparative analysis of two rates, *Statistics in Medicine*, 4, 213-226, 1985.

[10] Newcombe, R. G., *Confidence Intervals for Proportions and Related Measures of Effect Size*, Chapman & Hall/CRC, 2012.

[11] Newcombe, R. G., Interval estimation for the difference between independent proportions:Conparison of eleven methods, *Statistics in Medicine*, 17, 873-890, 1998.

[12] Newcombe, R. G. and Nurminen, M., In Defence of Score Intervals for Proportions and their Differences, *Communications in Statistics – Theory and Methods*, 40, 7, 1271-1282, 2011.

[13] Santner. T.J. et al., Small-sample comparisons of confidence intervals for the difference of two independent binomial proportions, *Computational Statistics & Data Analysis*, 51, 5791–5799, 2007.

[14] Wilson, E. B., Probable inference, the law of succession, and statistical inference, *Journal of the American Statistical Association*, 22, 209–212, 1927.

[15] Woodward, *Epidemiology* : *Study Design and Data Analysis*, Chapman & Hall/CRC, 2004.

[16] Yoo. D.H. et al., A randomised, double-blind, parallel-group study to demonstrate equivalence in efficacy and safety of CT-P13 compared with innovator infliximab when coadministered with methotrexate in patients with active rheumatoid arthritis: the PLANETRA study. *Annals of the Rheumatic Diseases*, 72, 1613–1620, 2013.

[17] 飯塚政人，浜田知久馬．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 論文集, 461-473, 2013.

MCMCプロシジャを用いたNormalized Power Priorの実用的な実装

武田 純
アステラス製薬株式会社 開発本部 データサイエンス部

Practical Implementation of Normalized Power Prior with MCMC Procedure

Jun Takeda
Data Science, Global Development, Astellas Pharma Inc.

要旨

Bayes 流モデリングの基本は，パラメータの事前分布をデータで更新して事後分布を得ることである。その派生として事前分布を「過去のデータ」で更新し，更新された分布を新たに事前分布とみなして「現在のデータ」で更新することも考えられる。その際「過去のデータ」の尤度関数を 1 未満のべき数を用いてべき乗することにより，その重みを弱めることを意図して導出した事前分布を Power Prior と呼ぶ。Power Prior のべき数の部分を定数ではなく確率変数とみなしたモデルも存在するが，定数の場合のモデルからの単純な拡張はべき数の事後分布が過小な値をとる傾向にある。そこである値による基準化も施した Normalized Power Prior (もしくは Modified Power Prior) が提案されている。Normalized Power Prior の実装上の問題点は，この基準化のための値の算出に積分計算を伴うこと，及びその値を MCMC サンプラーに組み込むことにある。

本論文では，MCMC プロシジャを用いて Normalized Power Prior を実現する実用的な方法を提示する。積分計算は MCMC プロシジャの外側でモンテカルロシミュレーションにより実装され，求められた値は FCMP プロシジャにより定義された関数を通して MCMC プロシジャ内から参照される。一般的な方法論とともに 2 項分布モデルのもとでの具体的な実装方針も示し，その事後分布から Normalized Power Prior の性質を考察する。

キーワード：Bayesian modeling，historical data，Normalized Power Prior，Modified Power Prior，MCMC プロシジャ，FCMP プロシジャ

1. はじめに

Bayes 流モデリング基本は，パラメータの事前分布をデータで更新して事後分布を得ることである。その派生として，更新された事後分布を新たな事前分布とみなし，さらに別のデータで更新するといった手順の繰り返しも考えられる。一連の過程は Bayes 更新とも呼ばれ，パラメータの最終的な事後分布は最初の事前分布とそれぞれのデータの尤度の積に比例する。この枠組みの特徴は，データごとに特に重みづけを行わないことである。仮に得られるデータの順番が異なったとしても，最終的な事後分布は変わらない。

データが過去，現在と2段階で得られた状況を考える。これらのデータにBayes更新を当てはめてもよいが，過去より現在のデータに重きを置く，言い換えれば情報量を割り引いたもとで過去のデータの情報を用いて推論を行いたい場合も想定される。また事前分布を現在のデータで更新する枠組みに，情報量を割り引いたもとで過去のデータを活用して推論の精度を高めたい場合もあるであろう。

過去のデータの情報量を割り引く方法として，Power Prior ([5]など) が提案されている。Power Priorとは，現在のデータ (current data) から見た事前分布を，過去のデータを得る前の事前分布 (initial prior と呼ばれる) と1未満のべき数でべき乗した過去のデータ (historical data) の積により定義される。過去のデータの情報量は，べき数によりコントロールされる。1に近ければ過去のデータを最大限に活用し，0に近ければ過去のデータはパラメータの事後分布にほとんど影響を与えなくなる。

Power Priorを用いる際には，べき数を定める必要がある。べき数の定め方としては，専門家の意見[9]に基づいた定数とするほか，べき数そのものを確率変数とし，そこに事前分布を与えることも提案されている ([5]など)。

本論文ではべき数を確率変数として扱うPower Priorの1つである，Normalized Power Prior (Modified Power Priorとも呼ばれる，[2][3][9][10]) に注目し，SASによる実装の方法論を提示する。2節では他のPower Priorとの関連も含めて，Normalized Power Priorの理論を解説する。Normalized Power Priorを用いる際の問題点は，必要とされるある種の積分計算であり，3節にてSAS内でその積分計算を行う実用的な方法論を示す。4節では2項分布モデルのもとで，Normalized Power Priorを用いた事後分布の計算結果を他のPower Priorと比較して示し，5節でまとめる。

2 Power Priorの概要

本節ではPower Priorの数理についてまとめる。まずべき数を定数としたもとでのPower Priorである条件付きPower Priorを示し，その派生としてIbrahim and Chen型Power Prior及びNormalized Power Priorについて示す。なお多く用語や表記は文献[9]を参考にしている。簡単のため，以降の議論において事前分布は全てproper (積分が発散しない) であると仮定する。

2.1 条件付きPower Prior

$\boldsymbol{D}_0$ を過去のデータ (確率変数の実現値としての値のほか，共変量も含む)，$\boldsymbol{D}$ を $\boldsymbol{D}_0$ と同じ構造を持つ現在のデータ，$\boldsymbol{\theta}$ を推論の対象となるパラメータのベクトル，$\pi_0(\boldsymbol{\theta})$ をパラメータベクトルに対する過去のデータを得る前の事前分布としてもとでの，べき乗パラメータ a_0 (とりうる値は $0 \le a_0 \le 1$) を定数としたPower Prior $\pi(\boldsymbol{\theta}|\boldsymbol{D}_0, a_0)$ は以下に定義される ([5]など):

$$\pi_C(\boldsymbol{\theta}|\boldsymbol{D}_0, a_0) \propto L(\boldsymbol{D}_0|\boldsymbol{\theta})^{a_0}\pi_0(\boldsymbol{\theta}) \tag{1}$$

ここに $L(\boldsymbol{D}_0|\boldsymbol{\theta})$ はパラメータ $\boldsymbol{\theta}$ のデータ $\boldsymbol{D}_0$ に関する尤度とする。以上で定義された $\pi_C(\boldsymbol{\theta}|\boldsymbol{D}_0, a_0)$ が現在のデータ $\boldsymbol{D}$ に対する事前分布となる。$a_0 = 1$ ならば (1) は $\boldsymbol{D}_0$ の情報を最大限に利用することになり，$\pi_0(\boldsymbol{\theta})$ のデータ $\boldsymbol{D}_0$ 及び $\boldsymbol{D}$ による通常のBayes更新に対応する。また $a_0 = 0$ ならば，$\boldsymbol{D}_0$の情報は一切用

いられず，$\pi_0(\boldsymbol{\theta})$ は $\boldsymbol{D}$ のみで更新される。a_0 が所与であることから，(1) を条件付き (Conditional) Power Prior と呼ぶことにする。(1) に基準化定数を含めた式は以下に示される:

$$\pi_{\mathrm{C}}(\boldsymbol{\theta}|\boldsymbol{D}_0, a_0) = \frac{L(\boldsymbol{D}_0|\boldsymbol{\theta})^{a_0}\pi_0(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} L(\boldsymbol{D}_0|\boldsymbol{\theta})^{a_0}\pi_0(\boldsymbol{\theta})\, d\boldsymbol{\theta}} = \frac{L(\boldsymbol{D}_0|\boldsymbol{\theta})^{a_0}\pi_0(\boldsymbol{\theta})}{g(a_0|\boldsymbol{D}_0)} \tag{2}$$

ここに $g(a_0|\boldsymbol{D}_0) = \int_{\boldsymbol{\theta}} L(\boldsymbol{D}_0|\boldsymbol{\theta})^{a_0}\pi_0(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ である。

a_0 は「専門家の意見」によって決められる ([9]) ほか，異なる a_0 から構成された条件付き Power Prior の中から，事後分布に関する偏差情報量基準 (Deviance Information Criterion, DIC) が最も小さいものを選ぶ，すなわちモデル選択により定める方法論も存在する ([6])。

2.2 Ibrahim-Chen 型 Power Prior

Bayes 流の考え方をより追求するのなら a_0 も確率変数として取り扱うことになる。なおその場合，a_0 自体を直接定める必要はなくなるが，a_0 の事前分布とその定数パラメータは決める必要がある。a_0 に事前分布を与えることによる (1) からの最も素直な修正は，以下に示される式であろう ([5]など):

$$\pi_{\mathrm{IC}}(\boldsymbol{\theta}, a_0|\boldsymbol{D}_0) \propto L(\boldsymbol{D}_0|\boldsymbol{\theta})^{a_0}\pi_0(\boldsymbol{\theta})\,\pi(a_0) \tag{3}$$

ここに $\pi(a_0)$ は a_0 に関する事前分布であり，$0 \le a_0 \le 1$ であることから通常 β 分布が想定される。以降は (3) を Ibrahim-Chen 型 Power Prior と呼ぶことにする。(3) に基準化定数を含めた式は以下に示される:

$$\pi_{\mathrm{IC}}(\boldsymbol{\theta}, a_0|\boldsymbol{D}_0) = \frac{L(\boldsymbol{D}_0|\boldsymbol{\theta})^{a_0}\pi_0(\boldsymbol{\theta})\pi(a_0)}{\int_0^1\int_{\boldsymbol{\theta}} L(\boldsymbol{D}_0|\boldsymbol{\theta})^{a_0}\pi_0(\boldsymbol{\theta})\pi(a_0)\, d\boldsymbol{\theta}\, da_0} \tag{4-1}$$

$$= \frac{L(\boldsymbol{D}_0|\boldsymbol{\theta})^{a_0}\pi_0(\boldsymbol{\theta})\pi(a_0)}{\int_0^1 g(a_0|\boldsymbol{D}_0)\pi(a_0)\, da_0} \tag{4-2}$$

Ibrahim-Chen 型 Power Prior は比較的広く用いられているが，いくつかの研究で a_0 の事後分布が小さな値を取りやすいことが指摘されている ([2][3][10])。つまり a_0 の事後分布が 0 付近に集まりやすく，過去のデータを推論に活用する枠組みであるにもかかわらず，実際には過去のデータが $\boldsymbol{\theta}$ の事後分布にほとんど影響を与えていないという問題点が挙げられてきた。

更に Ibrahim-Chen 型 Power Prior は $L(\boldsymbol{D}_0|\boldsymbol{\theta})$ に定数を乗じた際に，分布が変わるという特徴を持つ。例えば 2 項分布モデルを考える。試行数を n，成功確率を p，実際に得られた成功数を r とした場合，2 項分布の確率関数には $n!/(r!\,(n-r)!)\cdot p^r(1-p)^{n-r}$ となる。そこでは $L(\boldsymbol{D}_0|\boldsymbol{\theta})$ に 2 項係数 $n!/(r!\,(n-r)!)$ を含めるか否かによって異なる $\pi_{\mathrm{IC}}(\boldsymbol{\theta}|\boldsymbol{D}_0, a_0)$ が得られる。すなわち$(n!/(r!\,(n-r)!))^{a_0}$ が乗じられるか否かの違いが生じる。この不定性，言い換えれば尤度原理を満たさない点も Ibrahim-Chen 型 Power Prior の欠点と見られることがある ([2])。

2.3 Normalized Power Prior

(3) は (1) に直接 a_0 を確率変数として取り扱う修正を施したものであるが，一旦基準化 (normalization) された式 (2) に対して a_0 を確率変数として取り扱う修正を施したものが，以下の Normalized Power Prior (Modified Power Prior とも呼ばれる) である ([2][3][10]):

$$\pi_{\mathrm{NPP}}(\boldsymbol{\theta}, a_0|\boldsymbol{D}_0) = \frac{L(\boldsymbol{D}_0|\boldsymbol{\theta})^{a_0}\pi_0(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} L(\boldsymbol{D}_0|\boldsymbol{\theta})^{a_0}\pi_0(\boldsymbol{\theta})\,d\boldsymbol{\theta}}\,\pi(a_0) = \frac{L(\boldsymbol{D}_0|\boldsymbol{\theta})^{a_0}\pi_0(\boldsymbol{\theta})}{g(a_0|\boldsymbol{D}_0)}\,\pi(a_0) \tag{5}$$

Normalized Power Prior は以下の良い性質を持つ:

1. Ibrahim-Chen 型ほど a_0 の事後分布が小さい値を取る傾向にない。
2. Ibrahim-Chen 型とは異なり $L(\boldsymbol{D}_0|\boldsymbol{\theta})$ に定数を乗じても分布は不変である。

一方 Normalized Power Prior を用いる場合，新たに以下の問題が生じる:

1. MCMC サンプラー内で a_0 が更新されるたびに $g(a_0|\boldsymbol{D}_0)$ を求める必要がある。a_0 更新の度に数値積分を行うのでは，計算が非常に高負荷になるという問題を生じる ([4][9])。
2. $\boldsymbol{\theta}$ が多次元の場合，数値積分は 1 次元の場合に比べて格段に難しくなる。

3.4 節ではモンテカルロシミュレーションによる積分とその積分結果の線形補完により，これらの問題を回避する方法を示す。

3. MCMC プロシジャによる Power Prior の実装

以下，各種 Power Prior について，一般的な実装法と，データが 2 項分布に従う場合の具体的な例について示す。

2 項分布モデルの例

ここでは 3 節における Power Prior 実装の例示と 4 節の数値例で用いている 2 項分布モデルを示す。現在のデータは $\boldsymbol{D} = \{y_1, \ldots, y_N, n_1, \ldots, n_N\}$ と表される。ここに n_i は i 番目のデータの組における試行の数であり，y_i はそのうち成功した数，N はそのようなデータの組の数である。過去のデータについても同様に，$\boldsymbol{D}_0 = \{y_{01}, \ldots, y_{0N_0}, n_{01}, \ldots, n_{0N_0}\}$ と表記することにする。パラメータベクトルを $\boldsymbol{\theta} = p$ (1 パラメータ) としたもとで，各確率分布を以下のように定める:

$$p \sim \mathrm{B}(1,1) \tag{6-1}$$
$$y_{0i} \sim \mathrm{Bin}(n_{0i}, p) \tag{6-2}$$
$$y_i \sim \mathrm{Bin}(n_i, p) \tag{6-3}$$
$$a_0 \sim \mathrm{B}(1,1) \tag{6-4}$$

ここに $\mathrm{B}(a,b)$ はパラメータ a，b を持つ β 分布，$\mathrm{Bin}(n,p)$ は成功確率 p のもとで n 回の試行に関する 2

項分布を表す。また $y_{01},\ldots,y_{0N_0},y_1,\ldots,y_N$ は互いに独立であるとする。(6-1) はパラメータの事前分布, (6-2) は過去のデータの生成に関する分布 (過去のデータの尤度), (6-3) は現在のデータの生成に関する分布 (現在のデータの尤度), (6-4) はべき数に関する事前分布に対応する。

3.1 データセットの構造

一般論

過去のデータと現在のデータを合わせて一つのデータセットとする。データセットは以下の変数を含む:

1. データ (確率変数ベクトルの実現値)。
2. データ (確率変数ベクトルに関わる広い意味での共変量)。通常の共変量のほか, 2 項分布の試行回数, 対数線形モデルにおけるオフセット項など, 推論の対象とならず値の与えられているパラメータも含む。
3. オブザベーションが過去のデータのものなのか現在のデータのものなのかを区別する変数。
4. オブザベーションのキー項目。MCMC プロシジャが複数のオブザベーションに対して処理を行う中で, 過去データに属するオブザベーションの中で特定の処理を 1 回のみ行うために使用。

1. と 2. は通常の解析で現れる変数である。Power Prior の実装にあたって, 特に 3. と 4. が追加されている。

2 項分布モデルの例

例えば $N_0 = 3,\ N = 2,\ \boldsymbol{D}_0 = \{y_{01} = 10, y_{02} = 5, y_{03} = 5, n_{01} = 25, n_{02} = 50, n_{03} = 25\},\ \boldsymbol{D} = \{y_1 = 5, y_2 = 15,\ n_{01} = 50, n_{02} = 50\}$ の場合, データセットの構成は表 1 のようになる。

表 1: 2 項分布モデルにおけるデータセット例

オブザベーション番号	HIST_CUR	REC_ID	Y	N
1	0	1	10	25
2	0	2	5	50
3	0	3	5	25
4	1	1	5	50
5	1	2	15	50

HIST_CUR はそれぞれのオブザベーションが過去もしくは現在のどちらに属するデータかを示し (HIST_CUR=0 が過去, HIST_CUR=1 が現在), REC_ID はオブザベーションのキー項目で, HIST_CUR と組み合わせることにより, オブザベーションが一意に特定される。なお 4 節の数値例においては過去, 現在それぞれ複数の実現値を 1 つにまとめた表 2 の形のデータセットを用いる。そこではデータの組に関する添え字が必要なくなるため, $\boldsymbol{D}_0 = \{y_0 = 20, n_0 = 100\},\ \boldsymbol{D} = \{y = 20, n = 100\}$ と表すことにする。

表 2: 2 項分布モデルにおけるデータセット例 (過去, 現在のデータをそれぞれ 1 レコードで表現)

オブザベーション番号	HIST_CUR	REC_ID	Y	N
1	0	1	20	100
2	1	1	20	100

3.2 条件付き Power Prior の実装

一般論

MCMC プロシジャにおいて条件付き Power Prior を実装する方法は，文献 [11] で示されている:

- model ステートメントは確率分布ではなく，プロシジャ内で指定された対数尤度を general 関数により呼び出す。
- データセットのオブザベーションごとの対数尤度は logpdf 関数や lpdfXXX 関数 (XXX は確率分布を特定するキーワード) を用いる，もしくはユーザー自身が対数尤度関数を書き下すことにより指定する。
- 過去のデータに対応するオブザベーションについては，対数尤度を a_0 倍する。
- a_0 はマルコフ連鎖やデータセットのオブザベーションに関して不変であるため，begincnst/endcnst ステートメント内で定数であることを明示しておく (必須ではないが，間違いを未然に防ぐ手段になりうる)。

2 項分布モデルの例

以下に過去，現在のデータがデータセット ALLDATA に 3.1 節に示されたフォーマットで格納されている場合のサンプルコードを示す ($a_0 = 0.3$とする):

```
proc mcmc data=ALLDATA;
  parms P 0.5;
  begincnst; A0=0.3; endcnst; /* 定数として a0 = 0.3 を指定 */
  prior P ~ beta(1,1);
  if (HIST_CUR=0) then do; LLIKE=A0*logpdf('binomial', Y, P, N); end;
  if (HIST_CUR=1) then do; LLIKE=logpdf('binomial', Y, P, N); end;
  model general(LLIKE);
run;
```

2 項分布の対数尤度に対して，過去のデータに対応するデータのみ重み A0 ($= a_0$) が与えられる。

3.3 Ibrahim-Chen 型 Power Prior の実装

一般論

条件付き Power Prior のコードにおいて，a_0 に定数ではなく事前分布を与えるように修正する。parms ステートメントで該当する変数が確率変数であることを明示し，prior ステートメントにより事前分布を指定する。

2 項分布モデルの例

以下にサンプルコードを示す:

```
proc mcmc data=ALLDATA;
  parms P 0.5;
  parms A0 0.5; /* この行を追加。begincnst/endcnstステートメントを削除 */
  prior P ~ beta(1,1);
```

```
prior A0 ~ beta(1,1); /* この行を追加。 */
if (HIST_CUR=0) then do; LLIKE=A0*logpdf('binomial', Y, P, N); end;
if (HIST_CUR=1) then do; LLIKE=logpdf('binomial', Y, P, N); end;
model general(LLIKE);
run;
```

なお「2 項係数を含まない Ibrahim-Chen 型 Power Prior (詳細は 4.2 節において記載) については，5 行目を以下に置き換えることになる:

```
if (HIST_CUR=0) then do; LLIKE=A0*(logpdf('binomial', X, P, N)-lcomb(N, X)); end;
```

binomial を引数とした logpdf 関数は 2 項係数を含んでいるため，そこから $\log(n!/(x!\,(n-x)!))$ を引く操作を行っている。

3.4 Normalized Power Prior の実装

一般論

以下 a)，b)，及び c) により実装する。

a) モンテカルロシミュレーションによる積分計算の近似

データステップにて以下の手順により，データセットを作成する。

1. 十分大きな M_θ に対して，$\pi_0(\boldsymbol{\theta})$ に従う独立な確率変数ベクトルの列 $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{M_\theta}$ をモンテカルロシミュレーションにより生成する。
2. 十分大きな M_a に対して，$a_{0,i} = i/M_a$ $(i = 0, \ldots, M_a)$ を定義し，近似により以下を得る:

$$g(a_{0,i}|\boldsymbol{D}_0) = \int_{\boldsymbol{\theta}} L(\boldsymbol{D}_0|\boldsymbol{\theta})^{a_{0,i}} \pi_0(\boldsymbol{\theta})\, d\boldsymbol{\theta} \approx \frac{1}{M_\theta} \sum_{j=1}^{M_\theta} L(\boldsymbol{D}_0|\boldsymbol{\theta}_j)^{a_{0,i}} \stackrel{\text{def}}{=} g_{0,i} \tag{7}$$

3. 第 i オブザベーションに $g_{0,i-1}$ が格納されるような $(M_a + 1)$ レコードを持つデータセットを作成する。

b) FCMP プロシジャによる関数の定義

以下に示す $g(\cdot|\boldsymbol{D}_0)$ を近似した関数 $\hat{g}(\cdot|\boldsymbol{D}_0)$ を FCMP プロシジャにより定義する:

- 関数の引数は a_0 $(0 \le a_0 \le 1)$ とする。
- read_array 関数を用いて，データセットに格納された $g_{0,0}, \ldots, g_{0,M_a}$ の値を FCMP プロシジャに渡す。
- 関数の返り値を線形補完により求める。すなわち $a_{0,i'} < a_0 < a_{0,i'+1}$ となる i' を特定したのち，$(a_0 - a_{0,i'}) \times (g_{0,i'+1} - g_{0,i'})/(a_{0,i'+1} - a_{0,i'}) + g_{0,i'}$ を返り値とする $(a_0 = a_{0,i'}$ の場合は $g_{0,i'}$ を返り値とする)。

c) MCMC プロシジャにおける Normalized Power Prior の実装

MCMC プロシジャ内では，Ibrahim-Chen 型 Power Prior 実装のためのコードに対して修正を施す。すなわち，対数尤度において $\log[\hat{g}(\cdot|\boldsymbol{D}_0)]$ の引き算が行われるようにする。$\hat{g}(\cdot|\boldsymbol{D}_0)$ は過去のデータ全体に対する尤度の基準化であることから，MCMC プロシジャに渡されるデータセットのレコードが複数であって

もこの引き算は 1 回のみ行われるように指定する。その際に 3.1 節で述べた，オブザベーションのキー項目が役に立つ。

2 項分布モデルの例

まず $L(\boldsymbol{D}_0|p)^{a_{0,i}}$ を求めるサンプルコードを以下に示す ($a_{0,i} = 0.01$のもとでの 1 つの p の実現値 ($= \boldsymbol{\theta}_j$, コード内では THETA_P) を生成するためのコード):

```
proc sort data= ALLDATA out=HISTDAT; by HIST_CUR REC_ID; where HIST_CUR=0; run;
data LIKE_A0;
  set HISTDAT; by HIST_CUR;
  retain LLIKE;
  A0=0.01; THETA_P=rand("beta", 1, 1);
  if first.HIST_CUR then do; LLIKE=0; end;
  LLIKE0=logpdf('binomial', X, THETA_P, N);
  LLIKE=LLIKE+LLIKE0;
  if last.HIST_CUR then do;
    LIKE_A0=exp(A0*LLIKE);
    output;
  end;
run;
```

途中の計算を尤度ではなく対数尤度を用いて行うことにより，数値計算上の不安定性の回避を試みている。

実際は以下の手順を踏む:

1. THETA_P をM_θ 回発生させる。
2. それぞれの THETA_P の値に対して，全ての $a_{0,i} = 0/M_a, 1/M_a, \ldots, (M_a - 1)/M_a, M_a/M_a$ のもとでの $L(\boldsymbol{D}_0|p)^{a_{0,i}}$ を計算する ($M_\theta \times (M_a + 1)$ 通りの計算が必要になる)。計算結果は別レコードにアウトプットする。
3. UNIVARIATE プロシジャなどで，共通の $a_{0,i}$ ごとの $L(\boldsymbol{D}_0|p)^{a_{0,i}}$ の平均値を得る。すなわち，$g_{0,i}\ (i = 0, \ldots, M_a)$ を得る。
4. $g_{0,i}$ を値として持つ 1 変数のデータセット MAT1 を作成する。レコード数は $(M_a + 1)$ となる。なお 3. において平均値以外の記述統計量も求めて複数の変数として MAT1 に格納するのなら，積分のシミュレーションによる近似に関する診断の一助になる。
5. MAT1 を $(M_a + 1) \times 1$ 行列と見立てて FCMP プロシジャから読み込む。

1. と 2. においては，複数の p の実現値，複数の $a_{0,i}$ における，過去のデータに関する複数のオブザベーションに対する処理となる。よってコンピュータのリソースが許すのなら SQL プロシジャなどでデカルト積の形のデータを作っておくことにより，比較的シンプルなプログラムを書くことができる。

以下は FCMP プロシジャによる関数 $\hat{g}(\cdot|\boldsymbol{D}_0)$ の実装のサンプルコードである (マクロ変数 M_A に M_a が格納されているものとする):

```
proc fcmp outlib=WORK.TEMP.G;
  function G(A0);
    array MAT2[%eval(&M_A.+1)1, 1] / nosymbols;
    RC=read_array("MAT1", MAT2);
    I_A0=A0*(&M_A.)+1;
    I_A0_R=ceil(A0*(&M_A.) +I);
    I_A0_L=floor(A0*(&M_A.)+1);
    Y_L=MAT2[I_A0_L, 1];
    Y_R=MAT2[I_A0_R, 1];
    if I_A0_L=I_A0_R then do; Y=Y_L; end;
    else do; Y=(I_A0-I_A0_L)*(Y_R-Y_L)/(I_A0_R-I_A0_L)+Y_L; end;
    return(Y);
  endsub;
run;
options cmplib = WORK.TEMP;
```

最後の options ステートメントにより，関数 G が使用可能になる。

以下が MCMC プロシジャの使用に関するサンプルとなる:

```
proc mcmc data=ALLDATA;
  parms P 0.5;
  parms A0 0.5;
  prior P ~ beta(1,1);
  prior A0 ~ beta(1,1);
  if (HIST_CUR=0) then do; LLIKE=A0*logpdf('binomial', Y, P, N); end;
  if (HIST_CUR=0 and REC_ID=1) then do; LLIKE=LLIKE-log(G(A0)); end; /* この行のみ追加 */
  if (HIST_CUR=1) then do; LLIKE=logpdf('binomial', Y, P, N); end;
  model general(LLIKE);
run;
```

Ibrahim-Chen 型 Power Prior のコードから 1 行のみ追加されている。線形補完による積分計算の近似は MCMC サンプラー内でも行うことは可能と思われるが，コードが煩雑になることは明らかなため，関数呼び出しを行う方法を採用した。

4. 数値例

本節では 3 節で提示した方法論について，2 項分布モデルのもとで例示する。

4.1 シナリオ及び MCMC の設定

例示には 3 節で実装方針を示した 2 項分布モデルを用いる。表 3 に例示で用いた仮想データを示した。シ

ナリオ 1-2，1-4，2-2，2-4 は文献 [10] で示されているものである。現象をより理解するため，更にシナリオ 1-1，1-3，2-1，2-3 を追加した。SAS コード実装における各種パラメータについては表 4 に示した。

表 3: 仮想データ

シナリオ	過去のデータ (x_0/n_0)	現在のデータ (x/n)	解釈
1-1	6/30	6/30	過去と現在のデータが同じ性質を持つもとでデータ量が異なる。
1-2	20/100	20/100	
1-3	60/300	60/300	
1-4	200/1000	200/1000	
2-1	3/30	6/30	過去と現在のデータが異なる性質を持つもとでデータ量が異なる。
2-2	10/100	20/100	
2-3	30/300	60/300	
2-4	100/1000	200/1000	

表 4: SAS コード実装における各種パラメータ

パラメータ	値
M_a: $\hat{g}(\cdot)$ の近似の細かさ	1000 ($a_{0,0} = 0.000,\ \ a_{0,1} = 0.001, \ldots, a_{0,999} = 0.999, a_{0,1000} = 1.000$)
M_θ: 積分を近似するシミュレーションの繰り返し数	100000
MCMC の burn-in の長さ	10000
MCMC のチェーンの長さ	100000 (burn-in を含まず)
MCMC における thinning	thinning は行わない

4.2 べき数を確率変数とした Power Prior の事後分布

本節では以下の 3 つの Power Prior を用いて例示を行う (<>内は略号):

- 2 項係数を含まない Ibrahim-Chen 型 Power Prior <ICN>。事前分布の尤度関数は，2 項係数を含まない以下となる:

$$L_1(\boldsymbol{D}_0|p) = p^{y_0}(1-p)^{n_0-y_0}$$

- 2 項係数を含む Ibrahim-Chen 型 Power Prior <ICY>。事前分布の尤度関数は 2 項係数を含む以下となる:

$$L_2(\boldsymbol{D}_0|p) = \frac{n_0!}{y_0!\,(n_0-y_0)!}\,p^{y_0}(1-p)^{n_0-y_0}$$

- Normalized Power Prior <NPP>。理論上 $L_1(\boldsymbol{D}_0|p)$，$L_2(\boldsymbol{D}_0|p)$ のいずれを用いても同じ分布となる。数値計算においては $L_2(\boldsymbol{D}_0|p)$ を用いた。

文献 [10] では ICN と NPP について例示がなされている。これらの Power Prior においては，基準化定数

を無視すれば a_0 の事後分布の周辺密度関数をΓ関数で表現することができる。よって現在のデータが得られた後の確率変数 a_0 に関して，一定の区間内の値を取りうる確率や，モーメントなどの計算を 1 次元の数値積分で行うことが可能となる。

表 5 に各シナリオについて 3 つの Power Prior を用いたもとでの a_0 の事後平均と 95%信用区間を示した。以下の傾向が読み取られる:

- いずれのシナリオにおいても事後平均は ICN，ICY，NPP の順に大きくなる。
- 過去と現在のデータが同じ性質を持つ場合 (シナリオ 1-1～1-4)，ICN は標本サイズが小さいもとでも a_0 の事後分布は 0 に近い値をとりがちであり，標本サイズが大きくなるにつれて分布は 0 に収束する。一方 NPP においては標本サイズに関わらず平均は 0.5 を上回り，信用区間幅もほぼ一定である。ICY は ICN と NPP の間の結果となった。
- 過去と現在のデータが異なる性質を持つ場合 (シナリオ 2-1～2-4)，標本サイズが大きくなるにつれて事後分布は 0 に収束する。ただし ICN は標本サイズが小さいもとでもすでに事後平均が小さい値である一方，NPP は標本サイズの増加に伴い事後平均が急速に 0 に接近する。ICY は ICN と NPP の間の結果となった。

表 5: a_0 を確率変数とした Power Prior のもとでの a_0 の事後平均及び 95%信用区間 (括弧内)

シナリオ	2 項係数を含まない Ibrahim-Chen 型 Power Prior (ICN)	2 項係数を含む Ibrahim-Chen 型 Power Prior (ICY)	Normalized Power Prior (NPP)
1-1	0.066 (0.002, 0.240)	0.340 (0.010, 0.925)	0.568 (0.059, 0.981)
1-2	0.020 (0.000, 0.074)	0.301 (0.010, 0.894)	0.571 (0.068, 0.980)
1-3	0.007 (0.000, 0.024)	0.269 (0.008, 0.849)	0.577 (0.072, 0.983)
1-4	0.002 (0.000, 0.007)	0.239 (0.006, 0.802)	0.579 (0.074, 0.980)
2-1	0.091 (0.002, 0.340)	0.303 (0.009, 0.906)	0.524 (0.046, 0.975)
2-2	0.028 (0.001, 0.102)	0.197 (0.005, 0.751)	0.431 (0.030, 0.960)
2-3	0.009 (0.000, 0.034)	0.090 (0.002, 0.378)	0.198 (0.011, 0.717)
2-4	0.003 (0.000, 0.010)	0.027 (0.001, 0.102)	0.045 (0.003, 0.148)

また表 6 において，p の事後平均と 95%信用区間を示した (なお文献 [10] では p の事後分布については特に触れられていない)。傾向を以下にまとめた:

- 標本サイズの小さいシナリオでは事後平均が 0.2 を超えているものがある。これは p の事前分布が $B(1,1)$ であるためであり (試行を 2 回行い 1 回成功した場合のデータと等価の情報量。文献 [8] などに解説あり)，その影響が過去，現在のデータに比べて打ち消されていないためである。
- 過去と現在のデータが同じ性質を持つ場合，事後平均は ICN，ICY，NPP 間で大きくは違わないが，信頼区間の幅は ICN，ICY，NPP の順に狭くなっていく。表 1 の結果に示される通り，この順に過去のデータの情報をより多く取り込むためである。
- 過去と現在のデータが異なる性質を持つ場合，標本サイズを大きくするにつれ，ICN は 0.2 より大きい値から 0.2 に近づいていくのに対し，ICY，NPP は 0.2 より小さい値から 0.2 に近づいていく。また

ICY より NPP の事後平均の値が小さい。これらは表 1 における a_0 の事後分布平均により示される，過去のデータの取り込みの度合の違いに起因する。a_0 が大きな値をとるシナリオと Power Prior の組み合わせほど過去のデータの影響をより受けるため，p の事後平均はより小さな値となる。ICN は過去のデータの影響よりも元の事前分布 $B(1,1)$ の影響のほうが大きいため，事後平均が 0.2 より大きくなったと思われる。Power Prior 間で a_0 の事後平均の差が顕著なシナリオ 2-1，2-2 では，p の信頼区間幅は ICN，ICY，NPP の順に狭くなっていく。一方，a_0 の事後分布が 0 に近づいているシナリオ 2-3，2-4 においては，Power Prior 間で p の信頼区間幅の違いはほとんどない。

表 6: a_0 を確率変数とした Power Prior のもとでの p の事後平均及び 95%信用区間 (括弧内)

シナリオ	2 項係数を含まない Ibrahim-Chen 型 Power Prior (ICN)	2 項係数を含む Ibrahim-Chen 型 Power Prior (ICY)	Normalized Power Prior (NPP)
1-1	0.217 (0.098, 0.367)	0.215 (0.103, 0.353)	0.213 (0.110, 0.339)
1-2	0.206 (0.134, 0.289)	0.205 (0.139, 0.280)	0.204 (0.144, 0.273)
1-3	0.202 (0.159, 0.249)	0.202 (0.162, 0.244)	0.201 (0.166, 0.240)
1-4	0.201 (0.176, 0.226)	0.200 (0.178, 0.224)	0.201 (0.181, 0.221)
2-1	0.210 (0.093, 0.361)	0.195 (0.088, 0.338)	0.182 (0.086, 0.312)
2-2	0.203 (0.131, 0.285)	0.191 (0.124, 0.271)	0.177 (0.116, 0.256)
2-3	0.201 (0.157, 0.248)	0.194 (0.151, 0.241)	0.187 (0.143, 0.234)
2-4	0.200 (0.176, 0.226)	0.198 (0.174, 0.224)	0.196 (0.172, 0.222)

表 1 における標本サイズの増加に伴う a_0 事後分布の形状の変化を視覚的に確認するため，図 1 に事後分布のカーネル密度推定の結果を示した (なお有限区間 (0, 1) 上の推定のため，端の推定が若干歪んでいる)。特に NPP において，シナリオ 1-1～1-4 において事後分布の形状はほとんど変わらないが，シナリオ 2-1～2-4 においては分布が左にシフトしていくことが分かる。

5. おわりに

本報告では Normalized Power Prior の実用的な実装法について，実例を交えて提示した。方法論は FCMP プロシジャでユーザーにより定義された関数を，MCMC プロシジャから呼び出すことが可能であることを活用した。関数呼び出しの度に数値積分を行うことを回避するため，予め指定された引数の値に対応する積分値を求めておき，関数呼び出しの際には求めてある積分値を線形補完することにより返り値を求めた。積分計算は，被積分関数の性質を利用しモンテカルロシミュレーションにより求めた。

数値例では，2 項分布モデルのもとで Normalized Power Prior の性質を示した。そのもとでのパラメータは比率に関する 1 変数のみで，またその積分範囲は (0, 1) と有限なため，提示された方法で比較的安定した結果が得られたと考えられる。しかしパラメータが複数，もしくは積分範囲が無限区間となった場合は，結果が不安定になる可能性もある。特にモンテカルロシミュレーションによる積分計算の部分については，周辺尤度の計算 (文献 [7] 13 章) と同様の問題に注意する必要がある。

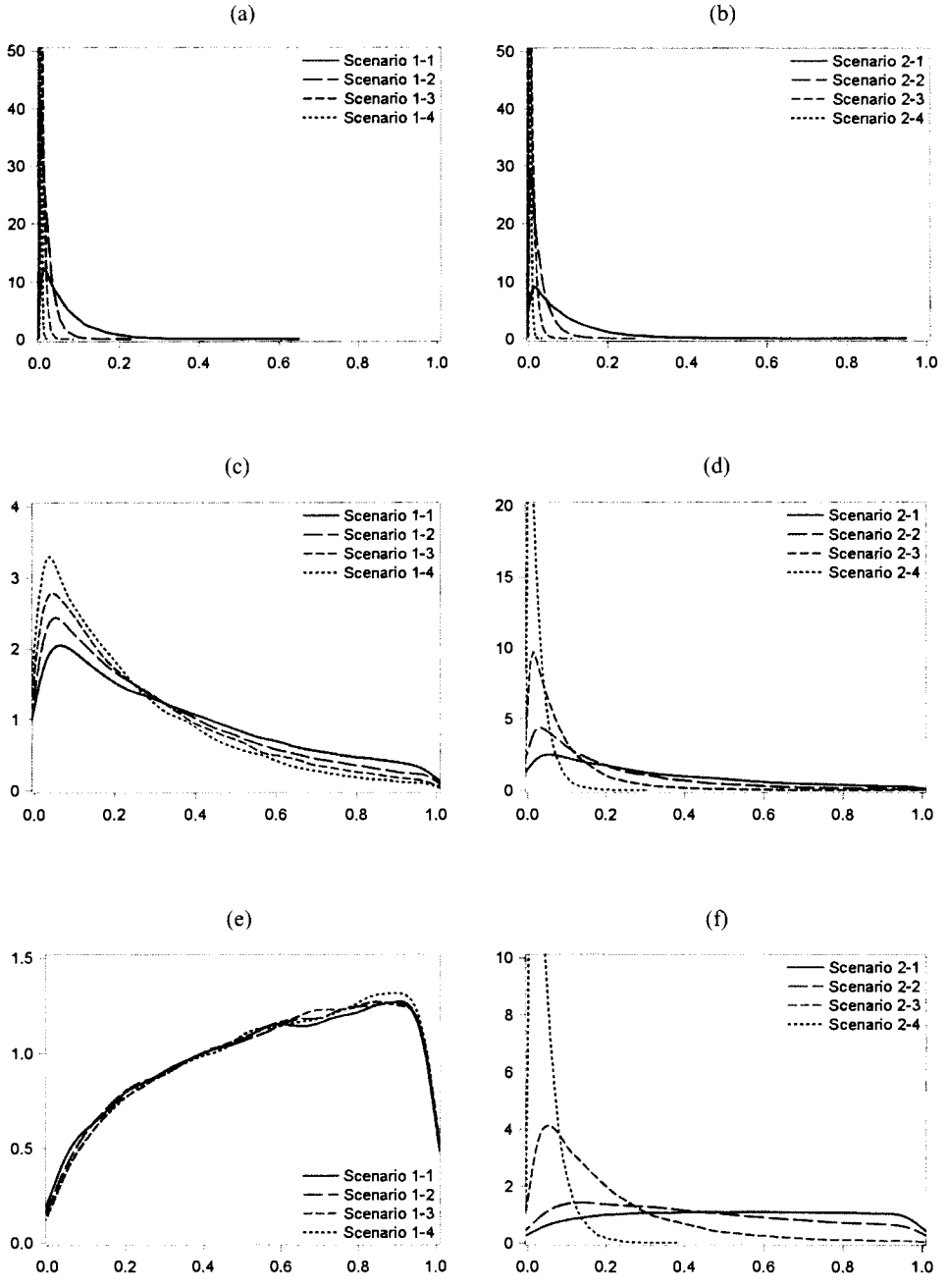


図 1 a_0 の事後分布のカーネル密度推定。(a), (b): 2 項係数を含まない Ibrahim-Chen 型 Power Prior (ICN)。(c), (d): 2 項係数を含む Ibrahim-Chen 型 Power Prior (ICY)。(e), (f): Normalized Power Prior (NPP)。

2 項分布モデルの数値例では，過去と現在のデータの性質が同じ場合は，Normalized Power Prior が Ibrahim-Chen 型 Power Prior と比べて，過去のデータの情報をより活用することが示された。一方，過去と現在のデータの性質が異なる場合，それぞれの情報量が増えるにつれ，Normalized Power Prior も過去のデータの情報を活用しなくなる。言い換えれば，過去と現在のデータの違いがよりはっきりするのなら，過去の情報の活用を控えるという特徴である。このような，過去のデータに対するデータ適応的な重みづけは望ましい特徴と考えられる ([2][3][10])。

Normalized Power Prior を更に発展させさた Commensurate Power Prior も提案されているが ([1])，べき数に加えて新たなパラメータを導入しており，実装にはかなりの工夫が必要になると思われる。よって複雑なモデルにおいても Normalized Power Prior がデータ適応的な重みづけの性質を持ち，また計算結果の数値的安定性が受け入れられるのであれば，本論文で示した方法論は有益になると考えられる。

参考文献

[1] Hobbs BP, Carlin BP, Mandrekar SJ, and Sargent DJ. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, **67**, 1047-1056.

[2] Duan Y, Ye K and Smith EP. (2006). Evaluating water quality using power priors to incorporate historical information. E*nvironmetrics*, **17**, 95–106.

[3] Duan Y, Smith EP, and Ye K. (2006). Using power priors to improve the binomial test of water quality. *Journal of Agricultural, Biological and Environmental Statistics*, **11**, 151–168.

[4] Gajewski BJ. (2010). Comments on 'A note on the power prior' by Neuenschwander B, Branson M and Spiegelhalter DJ. *Statistics in Medicine*, **29**, 708-710.

[5] Ibrahim JG and Chen MH. (2000). Power prior distributions for regression models. *Statistical Science*, **15**, 46–60.

[6] Ibrahim JG, Chen MH, and Chu H. (2012). Bayesian methods in clinical trials: a Bayesian analysis of ECOG trials E1684 and E1690. *BMC Medical Research Methodology*, **12**, 183.

[7] 小西貞則，越智義道，大森裕浩 (2008). 計算機統計学の方法 –ブートストラップ・EM アルゴリズム・MCMC–. 朝倉書店.

[8] Lunn D, Jackson C, Best N, Thomas A, and Spiegelhalter D (2012). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Chapman and Hall.

[9] Neelon B and O'Malley AJ. (2010). Bayesian analysis using power priors with application to pediatric quality of care. *Journal of Biometrics & Biostatistics*, **1**, 103.

[10] Neuenschwander B, Branson M, and Spiegelhalter DJ. (2009). A note on the power prior. *Statistics in Medicine*, **28**, 3562-3566.

[11] SAS Institute Inc. (2012). Bayesian binomial model with power prior using the MCMC procedure. (2014/06/27 に SAS Institute Inc. のウェブサイトにて存在を確認).

多重補完法におけるPattern-Mixtureモデルに基づく感度分析

伊藤 陽一[1] 西本 尚樹[2]
1 北海道大学大学院医学研究科医学統計学分野
2 北海道科学大学保健医療学部

Sensitivity analysis with the pattern-mixture model approach to multiple imputation

Yoichi M. Ito1, Naoki Nishimoto2
1 Department of Biostatistics, University of Hokkaido Graduate School of Medicine
2 Department of Radiological Technology, Faculty of Health Sciences, Hokkaido University of Science

要旨

近年、臨床試験の主要評価項目の解析において、欠測値を考慮した解析が推奨されている。多重補完法(multiple imputation methods)は、欠測値を補完するモデルを用いて欠測値の補完を行い、欠測値が埋められた完全データを複数作成し、完全データに対して予定された解析を行い、得られた複数の結果を併合する手法である。平均値補完など、単一の値を埋めて解析する手法と比較して、バラツキの過少評価を防ぐことができるため、欠測値を伴ったデータに対する手法として主要な手法の一つとなっている。しかし、多重補完法では、欠測の補完にあたっては、対象者が観測された値に従って、欠測が起こっているとする Missing at Random(MAR)を仮定しており、対象者が欠測値そのものの値に従って欠測が起こっている(Missing Not at Random; MNAR))場合には結果にバイアスが入ることが知られている。このMNAR の状況下では、バイアスのない推定値を求めることはできないため、結果変数が欠測にどの程度の影響を受けているかを調べる感度解析が行われることが多い。Pattern-Mixture モデルは、この感度解析を行うために提案されているモデルである。SAS9.4 から MI プロシジャに MNAR ステートメントが追加され、Pattern-Mixture モデルに基づいた感度解析が実行可能となった。本論文では、MI プロシジャにおける MNAR ステートメントの使用方法について解説を行う。

キーワード：多重補完法, Pattern-Mixture モデル, MI プロシジャ, MNAR ステートメント

はじめに

近年、臨床試験の主要評価項目の解析においては、欠測データが臨床試験の結果に大きく影響を与えることが指摘され、欠測値を考慮した解析が推奨されている[1)]。米国食品医薬品局（Food and Drug Administration; FDA）が、2008 年に全米研究評議会（National Research Council; NRC）の専門家パネルに対して依頼して作成された"The prevention and treatment of missing data in clinical trials"という報告書においても、欠測データの発生に関する仮定に対する感度解析を行うことが推奨されている[2)]。

欠測の仮定

欠測の仮定は、欠測が共変量(X)、補助変数(V)、結果変数(Y)のいずれにも依存しないMCAR (Missing Completely at Random)、観測された値のみに依存するMAR (Missing at Random)、欠測した値にも依存するMNAR (Missing Not at Random)に分類される。条件付分布の記法を用いると、MNARに関する分布がMARやMCARのときに、下記のように簡略化されることが分かる。ここで、V_{obs}は観測された補助変数、V_{mis}は欠測した補助変数、Y_{obs}は観測された結果変数、Y_{mis}は欠測した結果変数を表す。

$$\text{MNAR:}\quad [M|X,V_{obs},V_{mis},Y_{obs},Y_{mis}]$$
$$\text{MAR:}\quad [M|X,V_{obs},V_{mis},Y_{obs},Y_{mis}] = [M|X,V_{obs},Y_{obs},]$$
$$\text{MCAR:}\quad [M|X,V_{obs},V_{mis},Y_{obs},Y_{mis}] = [M]$$

欠測値を埋める解析方法 (Single imputation, Multiple imputation)

欠測値に対して単一の値で補完するSingle imputationのうち経時的な測定を伴う臨床試験で、広く用いられているのが、LOCF (Last Observation Carry Forward)法である。LOCF法は、経時的に観察される結果変数のうち途中で打ち切りになった症例については、最終観察値をもって、試験終了時点の観察値とするものである。アルツハイマー病などのQoL Scoreの自然悪化を伴う疾患に対する薬剤の治療目的は、悪化をできる限り遅らせることになるが、ある薬剤がプラセボと比較して早期に治療打ち切りが起こるとすると、LOCF では、薬剤の方がプラセボよりも、結果が良くなるというバイアスが入る (図1)[3]。

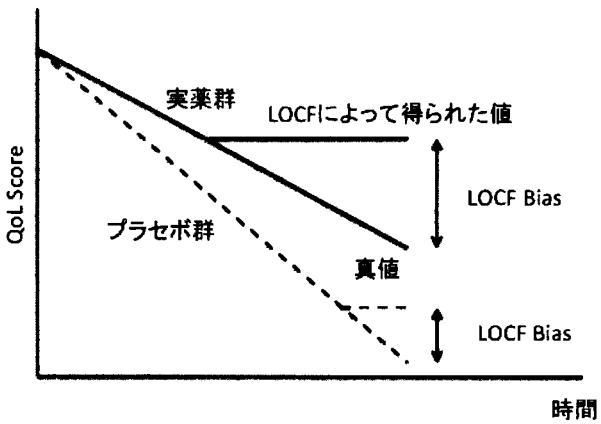


図1 LOCF Biasの概念図

このように、LOCFは、MCARの下であったとしても、バイアスを生じ得る。また、欠測を単一の値で補完するため、欠測に伴う不確実性について考慮できていない。したがって、LOCF は欠測に対する解析方法として、全く推奨できない。

LOCF以外のSingle imputationの方法としては、BOCF (Baseline Observation Carry Forward)法、regression imputation、hot deck imputationなどがある。BOCF法は、最終観察値ではなく、試験開始時点の値をもって、

試験終了時点の値する方法であるが、LOCF 法と同様に、バイアスが入り得る。Regression imputation 法は、観測されている値を用いて、欠測値を予測する回帰モデルを構築し、予測値によって欠測値を補完する方法である。Hot deck imputation 法は、欠測している結果変数以外の変数が類似している症例を選択し、その症例の結果変数をもって、結果変数の欠測を補完する方法である。どちらの方法も、欠測値を予測するプロセスが妥当であれば、LOCF 法や BOCF 法で生じるようなバイアスは減少させることができるが、欠測を単一の値で補完することによるバラツキの過少評価が生じる。

このバラツキの過少評価を改善する方法が、Multiple imputation 法である。Multiple imputation 法では、欠測値の予測分布からサンプリングを行い、欠測値を補完する。補完された完全データに対して通常の解析を実行する。その操作を複数回実行し、解析結果の平均を取る。この際に、欠測を補完したことに伴うバラツキについても考慮される。Multiple imputation 法の重要な利点は、最終解析に用いられない補助変数を補完モデルに用いることができる点である。多重補完法における予測分布は、MAR を仮定すると、ベイズ法における事後分布となる。

感度解析の原理と方法

これまでに紹介した解析は MAR を前提としてきた。主要な解析は MAR を前提とした解析を行い、観測データの MAR からの乖離は、MNAR を前提とした解析方法による感度解析で検討されるべきだと考える。解析の枠組みとして、2 つの仮定を考える。(i)欠測データの分布に関する検証不能な仮定と(ii)観測データの分布に関する検証可能な仮定である。完全データの分布は以下のように分割できる。

$$[Y_{obs}, Y_{mis}, M|X] = [Y_{obs}, M|X] \times [Y_{mis}|Y_{obs}, M, X]$$

$[Y_{mis}|Y_{obs}, M, X]$が仮定(i)に関するもの、$[Y_{obs}, M|X]$が仮定(ii)に関するものである。

感度解析手法としては、Pattern Mixture Model アプローチと Selection Model アプローチが提案されている。SAS9.4 で追加された MI プロシジャにおける MNAR ステートメントによって、Pattern Mixture Model アプローチに基づく感度解析を行うことができるようになった。Pattern Mixture Model の考え方は、欠測のパターンが先にあり、そのパターンごとにY_{mis}の分布が異なるというものである。条件付き分布の記法を用いると以下のように記述できる。

Pattern Mixture Model

$$[Y_{obs}, Y_{mis}, M|X] = [Y_{obs}, Y_{mis}|M, X] \times [M|X]$$
$$[Y_{obs}, Y_{mis}, M|X] = [Y_{mis}|Y_{obs}, M, X] \times [Y_{obs}|M, X] \times [M|X]$$

例として、結果変数が 1 つで、補助変数がない場合について示す。
Pattern Mixture Model の場合、観測できるかできないかで、結果変数Yの期待値が異なる。

$$E(Y|R = 0) = E(Y|R = 1) + \Delta$$
$$\mu_0 = \mu_1 + \Delta$$

この期待値の違いを表すΔが感度解析パラメータである。より一般的には、関数$g()$を用いて、下記のように書ける。

$$\mu_0 = g^{-1}(g(\mu_1) + \Delta)$$

完全データにおける結果変数の期待値μは、欠測確率をπとすると以下のようになる。

$$\mu = \pi\mu_1 + (1-\pi)g^{-1}(g(\mu_1) + \Delta)$$

この式において、μ_1とπを観測データから得られた$\hat{\mu}_1$と$\hat{\pi}$に置き替え、Δごとにμを求めることによって、感度解析を行う。

SAS の MI プロシジャにおける Pattern-Mixture Model

SAS での簡単な解析例を示す。臨床試験において、ベースライン共変量Y_0と割付群Trt(1: 実薬 0: プラセボ)には欠測がないものとする。主要評価項目Y_1に対して、下記の解析を行うことを考える。

$$Y_1 = \mu + \beta_1 Trt + \beta_2 Y_0$$

このとき、β_1が薬剤の効果を推定していることになる。臨床試験においては、薬剤に効果がないという仮定のもと解析を行うことが自然なので、Y_1の欠測に対しては、プラセボ群のデータのみを用いて埋めることが妥当である。これはプラセボ群のデータのみを用いて、Pattern Mixture Model におけるΔ= 0の下での補完を考えていくことになる。欠測データの構造としては、以下のようになる。

Obs	Trt	y0	y1
1	0	10.5212	11.3604
2	0	8.5871	8.5178
3	0	9.3274	.
4	0	9.7519	.
5	0	9.3495	9.4369
6	1	11.5192	13.2344
7	1	10.7841	.
8	1	9.7717	10.9407
9	1	10.1455	10.8279
10	1	8.2463	9.6844

このとき、下記のプログラムによって、プラセボ群のみを用いた多重補完を行うことができる。

```
proc mi data=Mono1 seed=14823 nimpute=10 out=outex15;
    class Trt;
    monotone reg (/details);
    mnar model( y1 / modelobs= (Trt='0'));
    var y0 y1;
run;
```

MNAR ステートメントでは、単調な欠測か、欠測が他の変数の条件付き分布として特定できることを前提としているので、MONTONE ステートメントか FCS ステートメントを同時に指定する必要がある。MNAR ステートメントの model オプション内の modelobs によって、どの対象者を用いて補完するかを指定することができる。ここではTrt = 0のプラセボ群のみを用いて補完するよう指定している。また、補完に用いる変数はY_0とY_1のみである。

Method として、Monotone が指定されており、Monotone ステートメントで reg オプションが指定されているので、MI プロシジャは回帰モデルに基づいてY_1補完を行っている。

The MI Procedure

Model Information

Data Set	WORK.MONO1
Method	Monotone
Number of Imputations	10
Seed for random number generator	14823

Monotone Model Specification

Method	**Imputed Variables**
Regression	y1

Missing data pattern テーブルでは、それぞれの missing pattern の頻度が集計される。下記の例では、25%の欠測が生じていることが分かる。

Missing Data Patterns

Group	**y0**	**y1**	**Freq**	**Percent**	**Group Means**	
					y0	**y1**
1	X	X	75	75.00	9.996993	10.709706
2	X	.	25	25.00	10.181488	.

MNAR ステートメントで model オプションを指定しているため、以下の出力がなされる。

Observations Used for Imputation Models Under MNAR Assumption

Imputed Variable	**Observations**
y1	Trt = 0

Details オプションを指定しているため、補完に用いた回帰モデルの回帰係数が出力されている。

Regression Models for Monotone Method

ImputedEffect	Obs-Data	Imputation									
Variable		1	2	3	4	5	6	7	8	9	10
y1	Intercept	-0.30169	-0.17426	-0.28040	-0.27518	30.09060	1-0.45748	0-0.24190	9-0.50135	1-0.05846	0-0.43665	0-0.50994	9
y1	y0	0.69364	0.64173	3 0.62997	0 0.50777	60.75228	3 0.83100	1 0.97007	5 0.72458	4 0.62363	8 0.56349	9 0.62128	0

補完された結果は以下のようになり、MIANALYZE プロシジャによって解析可能なデータセットが作られる。

Obs	_Imputation_	_Trt	y0	y1
1	1	0	10.5212	11.3604
2	1	0	8.5871	8.5178
3	1	0	9.3274	9.5786
4	1	0	9.7519	9.6060
5	1	0	9.3495	9.4369
6	1	1	11.5192	13.2344
7	1	1	10.7841	10.7873
8	1	1	9.7717	10.9407
9	1	1	10.1455	10.8279
10	1	1	8.2463	9.6844

$\Delta \neq 0$の下での補完例については、講演にて紹介する。

参考文献

1. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med. 2012 Oct 4;367(14):1355-60. PubMed PMID: 23034025. Epub 2012/10/05. eng.

2. National Research Council (U.S.). Panel on Handling Missing Data in Clinical Trials., National Research Council (U.S.). Committee on National Statistics., National Academies Press (U.S.). The prevention and treatment of missing data in clinical trials. Washington, D.C.: National Academies Press; 2010. xv, 144 p. p.

3. O'Neill RT, Temple R. The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it. Clinical pharmacology and therapeutics. 2012 Mar;91(3):550-4. PubMed PMID: 22318615.

隠れマルコフモデルによる予測モデルの構築

稲葉 洋介/株式会社新日本科学

宮岡 悦良/東京理科大学

Constructing prediction model by Hidden Markov Model

Yosuke Inaba / Shin Nippon Biomedical Laboratories

Etuo Miyaoka / Tokyo university of science

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

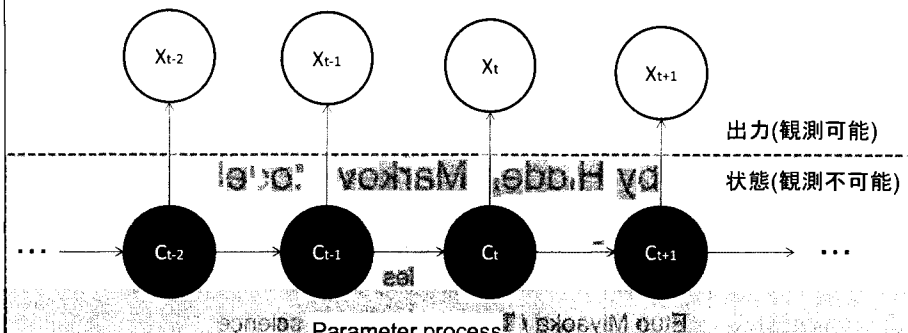
要旨:

隠れマルコフモデルを用いた予測モデルをSAS/IML を用いて実装し、実際のデータに適用した。

Forecasting state prediction hidden markov model flu

隠れマルコフモデルとは

- 観測不可能なマルコフチェーンの状態列(Parameter Process)、及び各状態をパラメータに持つ確率分布からの観測可能な出力により構成される確率モデル



定義：マルコフチェーン（1）

$C^{(t)} = \{C_t \in \{1, \ldots, m\} \mid t \in \mathbb{N}\}$:離散確率変数列

任意の $t \in \mathbb{N}$ に対して $C^{(t)}$ が以下を満たす時、$C^{(t)}$ をマルコフチェーンと呼ぶ。

$$\Pr(C_{t+1} \mid \mathbf{C^{(t)}}) = \Pr(C_{t+1} \mid C_t)$$

マルコフチェーンのパラメータ

$\gamma_{ij} := \Pr(C_{s+t} = j \mid C_s = i)$:推移確率

（特に γ_{ij} が s に依存せずに決まる時、このマルコフチェーンを homogenious と呼ぶ。）

$\Gamma_{ij} := \gamma_{ij}$：推移確率行列

定義：マルコフチェーン（2）

$\delta := \{\Pr(C_1 = 1), \Pr(C_1 = 2), ..., \Pr(C_1 = m)\}$：初期確率

※ $\mathbf{1} := (1, ..., 1)$ として $\delta\mathbf{1} = 1$

特に $\delta\gamma = \delta$ の時、このマルコフチェーンを定常的 (stationary) と呼ぶ。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

定義：ポアソン分布

定数 $\lambda > 0$ に対し、自然数を値にとる確率変数 X が以下を満たす時、X はポアソン分布に従うと言う。

$$\mathrm{p}(X) = \frac{\lambda^k e^{-\lambda}}{k!}$$

以後、平均 λ のポアソン分布を $Poisson(\lambda)$ と表記する。

性質：ポアソン分布は再生性を持つ。すなわち

$$X \sim Poisson(\lambda), Y \sim Poisson(\mu) \Rightarrow X + Y \sim Poisson(\lambda + \mu)$$

定義：ポアソン隠れマルコフモデル

C^t:マルコフ性を満たす ' Parameter Process '
X^t:C^t からの出力
この時、$\{C^t, X^t\}$ は以下を満たす時に隠れマルコフモデルと呼ぶ。

$$\Pr(C_t \mid C^{(t-1)}) = \Pr(C_t \mid C_{(t-1)}) \quad (1)$$

$$\Pr(X_t \mid X^{(t-1)}, C^{(t)}) = \Pr(X_t \mid C_{(t)}) \quad (2)$$

特に $C^{(t)}$ がポアソン分布に従う時、ポアソン隠れマルコフモデルと呼ぶ。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

ポアソン隠れマルコフモデルのパラメータ

$\boldsymbol{\lambda} := (\lambda_1, ..., \lambda_m)$	各状態のポアソン分布のパラメータ
$\boldsymbol{\Gamma_{ij}} := (\gamma_{ij})$	推移確率行列
$\boldsymbol{\delta} := (\delta_1, ..., \delta_m)$	初期確率

予測分布（forecast distribution）(1)

$h > 0$ として、

$$\begin{aligned}\Pr(\boldsymbol{X}_{T+h} = x \mid \boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)}) &= \frac{\Pr(\boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)}, X_{T+h} = x)}{\Pr(\boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)})} \\ &= \frac{\boldsymbol{\alpha}_T \boldsymbol{\Gamma}^h P(x) \mathbf{1}}{\boldsymbol{\alpha}_T \mathbf{1}}\end{aligned}$$

$\boldsymbol{\phi}_T := \boldsymbol{\alpha}_T / \boldsymbol{\alpha}_T \mathbf{1}'$ として、

$$\Pr(\boldsymbol{X}_{T+h} = x \mid \boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)}) = \boldsymbol{\phi} T \boldsymbol{\Gamma}^h P(x) \mathbf{1}'$$

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

予測分布（forecast distribution）(2)

従って、予測分布は

$$\Pr(\boldsymbol{X}_{T+h} = x \mid \boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)}) = \sum_{i=1}^{m} \xi_i(h) p_i(x)$$

ここで、ξ_i はベクトル $\boldsymbol{\phi}_T \boldsymbol{\Gamma}^h$ の i 番目の要素。

状態予測(state prediction)

$$\Pr(C_{T+h} = i \mid \boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)}) = \boldsymbol{\alpha}_T \boldsymbol{\Gamma}^h(\cdot, i)/L_T = \boldsymbol{\phi}_T \boldsymbol{\Gamma}^h(\cdot, i)$$

ただし、$\boldsymbol{A}(\cdot, i)$ は行列 $\boldsymbol{A}$ の i 番目の列を示す。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

Reference

- W Zucchini and L L.MacDonald(2009). Hidden Markov Models for Time Series, CRC Press.
- R Durbin, S Eddy, A Korgh, and G Mitchison(1998). Biological Sequence Analysis:Probabilic Models of Proteins and Nucleic Acids. Cambridge University Press.
- Baum, L.E., Petrie,T., Soules, G.and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann.Math. Statist.* 41, 164-171
- 小西　貞則,越智　義道,大森　裕浩.　計算統計学の方法 朝倉書店

ICLIFETESTプロシジャを用いた区間打切りデータの解析と既存プロシジャによる結果との比較

西本　尚樹、伊藤　陽一
北海道科学大学　保健医療学部　診療放射線学科
北海道大学　大学院　医学研究科　医学統計学分野

Interval censored data analysis and its attribute using ICLIFETEST PROCEDURE

Naoki Nishimoto[1], Yoichi M. Ito[2]
1. Department of Radiological Technology, Faculty of Health Sciences Hokkaido University of Science
2. Department of Biostatistics, University of Hokkaido Graduate School of Medicine

要旨

医薬品の臨床試験において、生存時間を推定する際に打ち切りの取扱方法によって結果にバイアスが入る可能性がある。生存時間解析において打ち切りは、右側打ち切りとして扱われる事が多いが、非致死性のアウトカムであった場合、左側打切り、もしくは区間打ち切りとなることがある。区間打ち切りとは、定期的に検査を受けていてある時点の検査が陽性であったとき、その前の時点の検査で陰性であった時点と、陽性となった時点のどこかでイベントが起こっているが、詳細なイベント発生時点が特定できない状態とされる。この現象は、悪性度の低い腫瘍や糖尿病の合併症である網膜症等、イベントの発生が捕らえづらい評価項目で起こる。

医学分野においては、LIFETEST プロシジャを用いて、ノンパラメトリックな生存時間の群間比較が行われているが、区間打切りのデータを解析するためには、LIFEREG プロシジャでパラメトリックな手法を使わざるを得なかった。

このほど SAS9.4 より、ICLIFETEST プロシジャが追加された。ICLIFETEST プロシジャは、区間打切りのデータに対して、ノンパラメトリックな推定を可能にするプロシジャである。これにより LIFEREG プロシジャで指定する Weibull 分布等のパラメトリックな分布を仮定することなく、生存時間の比較を行うことが可能となった。本論文では、ICLIFETEST プロシジャの使用方法と性質について解説する。

キーワード：生存時間解析、区間打切り、ICLIFETEST プロシジャ

1. 背景

医薬品の臨床試験において、生存時間を推定する際に打ち切りの取扱方法によって結果にバイアスが入る可能性がある。生存時間解析において打ち切りは、右側打ち切りとして扱われる事が多いが、糖尿病における網膜症など、非致死性のアウトカムで精密検査を受けて陽性と判定される場合など、左側打切り、もしくは区間打ち切りとなることがある[1)]。区間打ち切りとは、定期的に検査を受けていて、ある時点の検査が陽

性であったとき、その前の検査で陰性であった時点と、陽性となった時点のどこかでイベントが起こっているが、詳細なイベント発生時点が特定できない状態とされる。この現象は、悪性度の低い腫瘍や糖尿病の合併症である網膜症、思春期の開始時点等、イベントの発生がとらえづらい評価項目で起こる。また工学の領域では、飛行機の部品の破損、原子炉の圧力管、鉄道の枕木の破損等の定期的な点検によって発見されるイベントが該当する。区間打切りには、Current Status Data と呼ばれるデータも存在する。これは、ある個人や個体に対して観察期間中一度しか観察を行わない場合を指しており、被験者 i の観察時点を Ci とすると、区間が(0, Ci]または(Ci, ∞) のどちらかであるという状態を指す[2]。すなわち、被験者 i についてイベントの発生時間 Ti が Ti≦Ci であれば、(0, Ci] の区間となるのに対して、Ci<Ti であれば(Ci, ∞)となる。これは、実験動物にできた腫瘍を観察するのに解剖しなければいけない等が該当する。

医学領域においては、LIFETEST プロシジャを用いて、ノンパラメトリックな生存時間の群間比較が通常行われるが、左側打切り、区間打切りのデータを解析するためには、LIFEREG プロシジャでパラメトリックな手法を使わざるを得なかった。このほど SAS9.4 より、ICLIFETEST プロシジャが追加された。ICLIFETEST プロシジャは、区間打切りのデータに対して、ノンパラメトリックな推定を可能にするプロシジャである。これにより LIFEREG プロシジャで指定する Weibull 分布等のパラメトリックな分布を仮定することなく、生存時間の比較を行うことが可能となった。

2. 目的

本論文の目的は、ICLIFETEST プロシジャの使用方法について解説し、その特徴を明らかにすることとした。

3. 方法

3.1 区間打切りデータの定式化

i=1…n までの被験者 i が、ある定められた期間$0 = a_{i0} < a_{i1} < \cdots < a_{im} < \infty$に観察されたとする。被験者 i が時点$a_{i,j-1}$でイベントを起こさなかったとすると、被験者 i は次の時点a_{ij}で観察され、区間$(a_{i,j-1}, a_{ij}]$でイベントが起こったか否かを判定される。被験者 i の生存時間を Ti とすると、区間打切りは区間(Li, Ri]に対して、以下の関係を持って表現される。

$$L_i < T_i \leq R_i$$

これにより、$F_i(t)$を生存時間 Ti の累積分布関数として、N 人の独立した被験者から得られる尤度関数 L は、

$$L = \prod_{i=1}^{n} \left(F_i(L_i) - F_i(R_i)\right)$$

で表される。データ{(Li, Ri], i=1, …, n}から、生存曲線が S(t)=Pr(Ti>t)のもとで、非重複区間{(q1, p1], …, (qm, pm]}が導かれる。生存関数のノンパラメトリック最尤推定量(nonparametric maximum likelihood estimator、NPMLE)は、非重複区間(q1, p1], …, (qm, pm]は、より少ない時にのみ減少する。従って、jump probabilities は、これらの区間のみで推定する必要がある[3,4]。生存関数は、これらの非重複区間の一部また全部で減少し、非重複区間以外では、一定であることが仮定されている。応答変数である生存時間に対して打切りのメカニズムが独立であり、個々の被験者が最終的にはイベントを起こすと仮定すると、生存時間データ$\{T_i \in (L_i, R_i], j = 1, \ldots, m\}$の尤度は、疑似パラメータである$\theta\ j\{\theta_j = Pr(q_j < T \leq p_j),\ j = 1, \ldots, m\}$から

構成される。パラメーターベクトル$\theta = (\theta_1, \ldots, \theta_m)$は、$\sum_{j=1}^{m} \theta_j = 1$という制約の下で、尤度L(θ)を最大化するように推定される。

$$L(\theta) = \prod_{i=1}^{n} \left[\sum_{j=1}^{m} z_{ij} \theta_j \right]$$

ここで、z_{ij}は、$(q_j, p_j]$が$(L_i, R_i]$に含まれていれば1を、含まれていなければ0を示す。最尤法は、$\{\hat{\theta}_1, \ldots, \hat{\theta}_m\}$を推定し、生存関数のノンパラメトリック最尤推定量を決定する。

$$\hat{S}(t) = \begin{cases} 1 & t < q_1 \\ \prod_{k=j+1}^{j} \hat{\theta}_k & p_j \le t \le q_{j+1} \\ 0 & t \ge p_m \end{cases}$$

Petoは、制限付きニュートン・ラフソン法を使用することで、対数尤度の最大値を探索する方法を提唱した[3]。しかし、疑似パラメータが多い場合、最適化できない。また、ニュートン・ラフソン法では、グローバルな最大値が保証されない。Turnbullは、尤度関数の最大化は、self-consistency equationと等価であることを証明し、EMアルゴリズムを用いて解が得られることを示した[4,5]。

$$\theta_j = \frac{1}{n} \sum_{i=1}^{n} \frac{z_{ij}\theta_j}{\sum_{k=1}^{m} z_{ik}\theta_k}$$

3.2 ICLIFETESTプロシジャの文法とサンプルデータの解析

ICLIFETESTプロシジャの文法を以下に示す。TIMEステートメント以外は、他のプロシジャで使用されている記法とほぼ同様である。特徴的な記法は、TIMEステートメントで、引数となるには、区間の開始時間及び終了時間を指定する。開始時刻を欠測としてピリオドを指定した場合には、左側打切りとなり、いわゆるcurrent status dataと呼ばれる研究開始時点から一度のみの計測を行うデータとなる。また、終了時点を欠測としてピリオドを指定した場合には、右側打切りとなり、解析時点でイベントを起こしていない状態を表現する。開始時点、終了時点が共に欠測している場合には、その区間を欠測として扱う。

```
PROC ICLIFETEST <options> ;
    BY variables ;
    FREQ variable ;
    STRATA variables ;
    TEST variable </options> ;
    TIME (variable,variable ) ;
```

ICLIFETESTステートメントには、DATA=オプション以外に、IMPUTEオプションがある。INPUTEオプションは、欠測データの補完法の一つであるmultiple imputationに対して、標準誤差や一般化ログランク統計量に対する共分散行列を推定するために、seedなどの詳細を指定するものである。IMPUTE(SEED=XXXX)のように、XXXXの部分を指定する。デフォルトでは、ランダムに指定される。

また、生存関数の推定量を計算する方法は、上記で述べたように種々の方法が提案されており、ICLIFETESTプロシジャにおいても、推定方法を指定することができる。METHOD=オプションでTURNBULL、ICMまたはEMICM

のいずれかを指定することができる。TURNBULL を指定すると、Turnbull が論文の中で提案した方法で、計算する[4]。これは、EM アルゴリズムであり、TURNBULL の代わりに、METHOD=EM でも同じ結果が得られる。しかしながら、推定すべきパラメータの数がそれほど多くなくとも、EM アルゴリズムは解への収束の遅さが知られている。そこで、Groeneboom らが提案した ICM アルゴリズム(iterative convex minorant)と Wellner らが提案した EM-ICM アルゴリズム(EM iterative convex minorant)が、EM アルゴリズムをよりも NPMLE を計算するためにより efficient な方法でありため、デフォルトでは、EMICM で計算される[6,7]。

生存関数の推定に用いたデータには、Appendix 1 に示す Finkelstein らの論文で使用された乳がん切除術後の cosmetic deterioration のリスクをレトロスペクティブに比較したデータを用いた。放射線治療を受けた群(RT 群)と化学療法と放射線治療の両方を受けた群(RT+RCT 群)の生存時間のデータを用いた。被験者は、4 ヶ月から 6 ヶ月毎に追跡され、deterioration に至る区間打切りのデータが収集された。94 人のデータのうち、右側打切りは 38 人、残りの 56 人が区間打切りまたは左側打切りとなった。それぞれの治療群のデータを結合し、データセット BCS とした。

以下に、ICLIFETEST プロシジャを用いた SAS コードを示す。

```
ods graphics on;
ods html body='ICLIFETEST.html' path='C:¥Users ';
PROC ICLIFETEST PLOTS=(survival logsurv) DATA=BCS IMPUTE(SEED=1234);
   STRATA TRT;
   TIME (lTime, rTime);
RUN;
```

PLOTS=オプションでは、推定された生存関数のグラフ及び推定値の対数変換したものを負の値にしたもののグラフを表示するよう指定した。STRATA ステートメントでは、治療群 TRT(RT または RT+RCT)を指定した。TIME ステートメントでは、区間の開始と終了に lTime 及び rTime を指定した。

既存手法の結果を示すために、LIFEREG プロシジャを用いた区間打切りデータの解析方法を以下に示した。ICLIFETEST プロシジャでは、lTime に相当するデータが 0、すなわち開始時点の場合、自動的に左側打切りとして解析されるが、LIFEREG プロシジャでは、欠測としなければ左側打切りとして処理されない。そこで、新たに BCS2 というデータセットを作成し、lTime が 0 のデータを欠測値に置き換えた。
PLOBPLOT ステートメントで、カプランマイヤータイプのプロットを作成した。これは、累積確率分布が直線になるか否かをみて、確率分布の当てはまりを判定する。OUTPUT ステートメントでは、累積確率の推定値を BCS2_LROUT データセットに出力した。

```
DATA BCS2;
        SET BCS;
        IF lTime = 0 then lTime = .;
run;

ODS GRAPHICS ON;
PROC LIFEREG DATA=BCS2 ;
```

```
        CLASS TRT;
        MODEL(lTime, rTime) = TRT / DIST=Weibull;
        PROBPLOT PPOS=KM
        OUTPUT OUT=BCS2_LROUT CDF=PROB;
run;
```

4. 結果

4. 1ICLIFETEST プロシジャによる生存確率の推定値

The ICLIFETEST Procedure

Stratum 1: trt = RT

Nonparametric Survival Estimates

Time Interval		Probability Estimate Failure	Survival	Imputation Standard Error
0	4	0.0000	1.0000	0.0000
5	6	0.0463	0.9537	0.0354
7	7	0.0797	0.9203	0.0458
8	11	0.1684	0.8316	0.0580
12	24	0.2391	0.7609	0.0629
25	33	0.3318	0.6682	0.0706
34	38	0.4136	0.5864	0.0739
40	46	0.5344	0.4656	0.0758
48	Inf	1.0000	0.0000	0.0000

The ICLIFETEST Procedure

Stratum 2: trt = RT+RCT

Nonparametric Survival Estimates

Time Interval		Probability Estimate Failure	Survival	Imputation Standard Error
0	4	0.0000	1.0000	0.0000
5	5	0.0433	0.9567	0.0332
8	11	0.0866	0.9134	0.0413
12	16	0.1558	0.8442	0.0567
17	18	0.3012	0.6988	0.0764
19	19	0.4423	0.5577	0.0835
20	24	0.5580	0.4420	0.0771
25	30	0.6579	0.3421	0.0740
31	35	0.7288	0.2712	0.0674
36	44	0.8896	0.1104	0.0482
48	48	0.9448	0.0552	0.0352
60	Inf	1.0000	0.0000	0.0000

Quartile Estimates

Percentile	Point Estimate	95% Confidence Interval Transform	[Lower	Upper)
75	.	LOGLOG	.	.
50	40	LOGLOG	34	48
25	25	LOGLOG	8	34

Quartile Estimates

Percentile	Point Estimate	95% Confidence Interval Transform	[Lower	Upper)
75	36	LOGLOG	25	36
50	20	LOGLOG	19	25
25	17	LOGLOG	12	19

A	B
C	D

Table 1 ノンパラメトリックの生存確率の推定値及び四分位点の推定値と 95%信頼区間。(A)RT 群のノンパラメトリックの生存確率の推定値、(B)RT 群の四分位点の推定値、(C)RT+RCT 群のノンパラメトリックの生存確率の推定値、(D) RT+RCT 群の四分位点の推定値。区間打切りが生じているため、deterioration の起こる確率と起こさずに生存する確率のノンパラメトリック推定値は、非重複区間の集合として計算される。これらの推定値は、同一の区間において一定である。

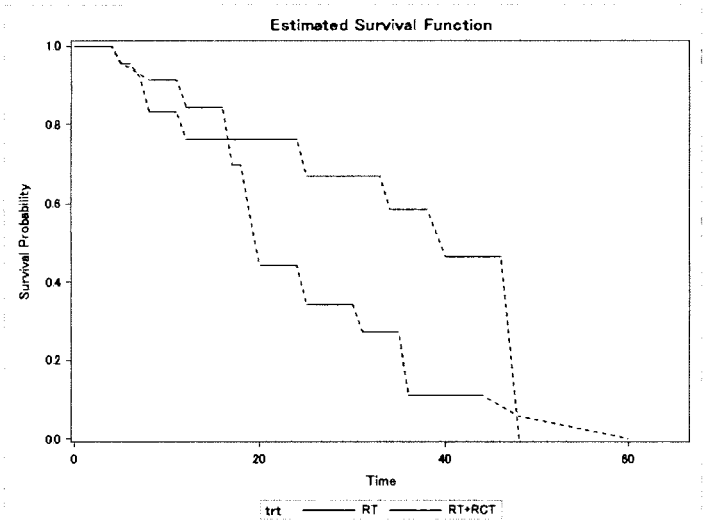


Fig. 1 区間打切りについて推定された生存関数。Table1(A), (B)で計算されたように、生存確率の推定値はTurnbull区間内では決定されない。従って、ICLIFETESTプロシジャでは、非重複区間の間を点線でつなぐ表示がなされる。NODASHオプションで非表示にすることも可能である。

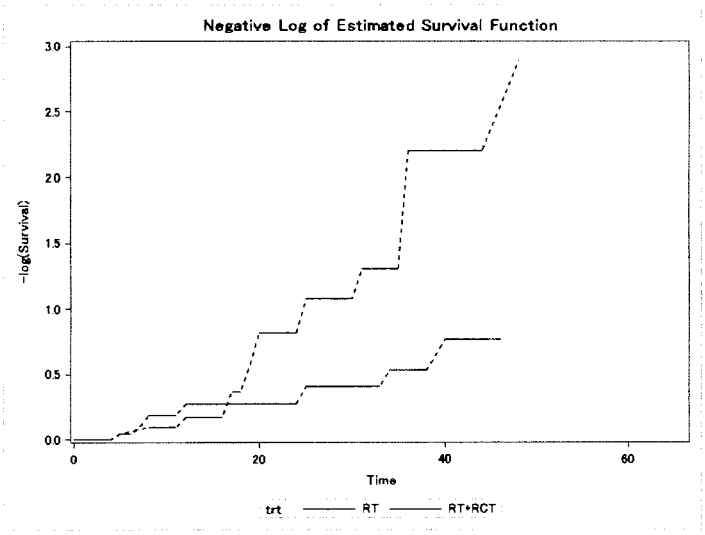


Fig. 2 負の対数変換を施したノンパラメトリック生存確率の推定値。

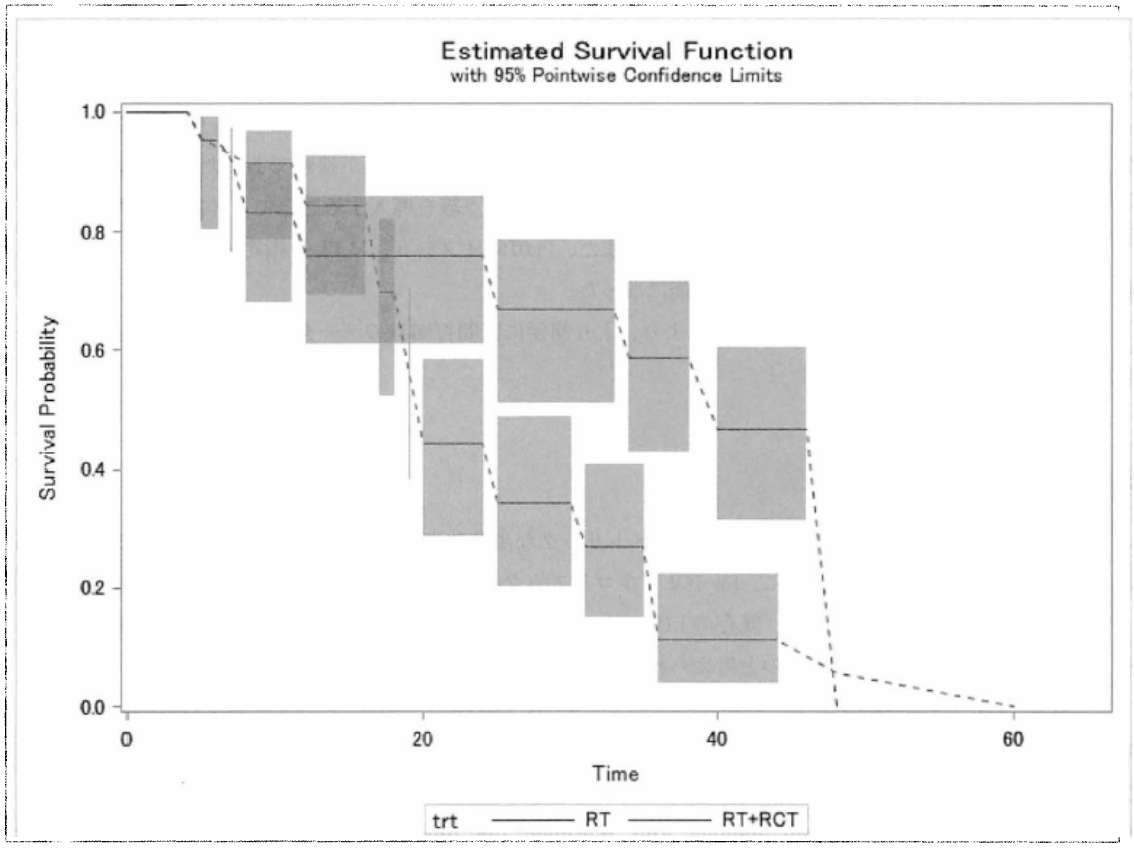


Fig.3 fig.1 に 95%信頼区間を追加した図。plots=オプションに、SURVIVAL(CL)と指定することで描画する。

4.2 LIFEREG プロシジャにより推定した累積発症率

Appendix 2 に LIFEREG プロシジャにより推定した累積発症率を示した。Weibull 分布などを当てはめることで、累積発症率を推定することはできるが、推定された区間打切りの累積発症率をグラフに表示する際に、累積発症率を示す時間が一意に決まらないため、推定値のみを示した。

5. 考察

これまでは、区間打切りのデータについては、LIFEREG プロシジャを用いて、Weibull 分布などの確率分布を当てはめて推定するか、SAS マクロを用いてノンパラメトリックな生存確率を推定する方法など、選択肢が限られていたが、プロシジャ化されたことにより、EM-ICM などの推定方法を簡便に扱うことができるようになった。

本論文では、ICLIFETEST プロシジャの使用方法を説明した。オプションが豊富に用意されており、最尤推定量の計算においても、3 種類の手法が利用可能である。Wellner らが提案した EM-ICM アルゴリズムが、すでにデフォルトで指定されており、NPMLE を推定するための高速な手法が利用可能である。また、各オプションの記法は、既存の生存時間を扱うプロシジャと一貫性があり、特に LIFEREG プロシジャの記法と非常によく似ているため、記法の再学習の必要性が軽減されている。ICLIFETEST プロシジャは、LIFEREG プロシジ

ャの記法に加えて、左側打切りの扱いが、0時点でも左側打切りとして処理されるなど、解析担当者にとってコード変換の負担が軽減するようになっている。

区間打切りのデータから生存確率を推定するためには、Turnbullが論文の中で述べている非重複区間の特定が必要になる。カプラン・マイヤー曲線の描画には、非重複区間の描画が課題であったが、Fig.1の(A)、(B)で示されるjump probabilityの推定が簡便になり、これも点線を導入することで、非重複区間をまたぐカプラン・マイヤー曲線の描画が可能になった。また、PLOTS=オプションにCLを指定することで、95%信頼区間の描画をODS経由で使用することが可能になった。

ICLIFETESTプロシジャが登場したことにより、より簡便に区間打切りのデータを扱うことが可能になるものと考えられる。

5. 結論

本論文では、区間打切りデータからノンパラメトリックな生存確率を推定するICLIFETESTプロシジャを解説した。生存確率を推定するために、EM-ICMアルゴリズムなどが実装されており、より高速に推定可能である。プロシジャ化されたことから、既存のLIFEREGプロシジャなどと、記法に一貫性をもち、区間打切りデータの解析及び研究に貢献するものと考えられる。

参考文献

1) 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.

2) Lawless JF. *Statistical models and methods for lifetime data*. 2nd ed. Hoboken, N.J.: Wiley-Interscience; 2003.

3) Peto R. Experimental Survival Curves for Interval-Censored Data. *Applied Statistics*. 1973;22:86-91.

4) Turnbull BW. The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data. *Journal of the Royal Statistical Society, Series B*. 1976;38:290–295.

5) Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorith. *Journal of the Royal Statistical Society, Series B*. 1977;39:1-38.

6) Groeneboom P, Wellner JA. *Information Bounds and Nonparametric Maximum Likelihood Estimation*. New York: Birkhauser; 1992.

7) Wellner JA, Zhan Y. A Hybrid Algorithm for Computation of the Nonparametric Maximum Likelihood Estimator from Censored Data. *Journal of the American Statistical Association*. 1997;92:945–959.

APPENDIX 1

ICLIFETESTプロシジャのexampleデータ。

```
data RT;
   input lTime rTime @@;
   trt = 'RT      ';
   datalines;
45   .   25 37   37   .
 6 10   46   .    0   5
```

```
 0   7   26 40   18   .
46   .   46   .   24   .
46   .   27 34   36   .
 7 16    36 44    5 11
17   .   46   .   19 35
 7 14    36 48   17 25
37 44    37   .   24   .
 0   8   40   .   32   .
 4 11    17 25   33   .
15   .   46   .   19 26
11 15    11 18   37   .
22   .   38   .   34   .
46   .    5 12   36   .
46   .
;

data RCT;
    input lTime rTime @@;
    trt = 'RT+RCT';
    datalines;
 8 12     0   5   30 34
 0 22     5   8   13   .
24 31    12 20   10 17
17 27    11   .    8 21
17 23    33 40    4   9
24 30    31   .   11   .
16 24    13 39   14 19
13   .   19 32    4   8
11 13    34   .   34   .
16 20    13   .   30 36
18 25    16 24   18 24
17 26    35   .   16 60
32   .   15 22   35 39
23   .   11 17   21   .
44 48    22 32   11 20
14 17    10 35   48   .
;
RUN;

data BCS;
    set RT RCT;
run;

PROC ICLIFETEST plots=(survival logsurv) data=BCS impute(seed=1234);
    STRATA TRT;
    TIME (lTime, rTime);
RUN;
```

Appendix 2 LIFEREG プロシジャによる累積発症率の推定値

RT 群

OBS	lTime	rTime	trt	_PROB_	PROB
1	45	.	RT	0.5	0.57731
2	25	37	RT	0.5	0.28347
3	37	.	RT	0.5	0.46622
4	6	10	RT	0.5	0.03273
5	46	.	RT	0.5	0.59025
6	.	5	RT	0.5	.
7	.	7	RT	0.5	.
8	26	40	RT	0.5	0.29892
9	18	.	RT	0.5	0.17809
10	46	.	RT	0.5	0.59025
11	46	.	RT	0.5	0.59025
12	24	.	RT	0.5	0.26808
13	46	.	RT	0.5	0.59025
14	27	34	RT	0.5	0.31439
15	36	.	RT	0.5	0.45152
16	7	16	RT	0.5	0.04179
17	36	44	RT	0.5	0.45152
18	5	11	RT	0.5	0.02449
19	17	.	RT	0.5	0.16376
20	46	.	RT	0.5	0.59025
21	19	35	RT	0.5	0.19266
22	7	14	RT	0.5	0.04179
23	36	48	RT	0.5	0.45152
24	17	25	RT	0.5	0.16376
25	37	44	RT	0.5	0.46622
26	37	.	RT	0.5	0.46622
27	24	.	RT	0.5	0.26808
28	.	8	RT	0.5	.
29	40	.	RT	0.5	0.50933
30	32	.	RT	0.5	0.39139
31	4	11	RT	0.5	0.01714
32	17	25	RT	0.5	0.16376
33	33	.	RT	0.5	0.40660
34	15	.	RT	0.5	0.13594
35	46	.	RT	0.5	0.59025
36	19	26	RT	0.5	0.19266
37	11	15	RT	0.5	0.08474
38	11	18	RT	0.5	0.08474
39	37	.	RT	0.5	0.46622
40	22	.	RT	0.5	0.23752
41	38	.	RT	0.5	0.48076
42	34	.	RT	0.5	0.42169
43	46	.	RT	0.5	0.59025
44	5	12	RT	0.5	0.02449
45	36	.	RT	0.5	0.45152
46	46	.	RT	0.5	0.59025

RT+RCT 群

OBS	lTime	rTime	trt	_PROB_	PROB
47	8	12	RT+RCT	0.5	0.12400
48	.	5	RT+RCT	0.5	.
49	30	34	RT+RCT	0.5	0.57330
50	.	22	RT+RCT	0.5	.
51	5	8	RT+RCT	0.5	0.06010
52	13	.	RT+RCT	0.5	0.25170
53	24	31	RT+RCT	0.5	0.54172
54	12	20	RT+RCT	0.5	0.22493
55	10	17	RT+RCT	0.5	0.17289
56	17	27	RT+RCT	0.5	0.36054
57	11	.	RT+RCT	0.5	0.19860
58	8	21	RT+RCT	0.5	0.12400
59	17	23	RT+RCT	0.5	0.36054
60	33	40	RT+RCT	0.5	0.72878
61	4	9	RT+RCT	0.5	0.04231
62	24	30	RT+RCT	0.5	0.54172
63	31	.	RT+RCT	0.5	0.69258
64	11	.	RT+RCT	0.5	0.19860
65	16	24	RT+RCT	0.5	0.33331
66	13	39	RT+RCT	0.5	0.25170
67	14	19	RT+RCT	0.5	0.27877
68	13	.	RT+RCT	0.5	0.25170
69	19	32	RT+RCT	0.5	0.41438
70	4	8	RT+RCT	0.5	0.04231
71	11	13	RT+RCT	0.5	0.19860
72	34	.	RT+RCT	0.5	0.74570
73	34	.	RT+RCT	0.5	0.74570
74	16	20	RT+RCT	0.5	0.33331
75	13	.	RT+RCT	0.5	0.25170
76	30	36	RT+RCT	0.5	0.57330
77	18	25	RT+RCT	0.5	0.38759
78	16	24	RT+RCT	0.5	0.33331
79	18	24	RT+RCT	0.5	0.38759
80	17	26	RT+RCT	0.5	0.36054
81	35	.	RT+RCT	0.5	0.76185
82	16	60	RT+RCT	0.5	0.33331
83	32	.	RT+RCT	0.5	0.71107
84	15	22	RT+RCT	0.5	0.30602
85	35	39	RT+RCT	0.5	0.76185
86	23	.	RT+RCT	0.5	0.51734
87	11	17	RT+RCT	0.5	0.19860
88	21	.	RT+RCT	0.5	0.46684
89	44	48	RT+RCT	0.5	0.87460
90	22	32	RT+RCT	0.5	0.49237
91	11	20	RT+RCT	0.5	0.19860
92	14	17	RT+RCT	0.5	0.27877
93	10	35	RT+RCT	0.5	0.17289
94	48	.	RT+RCT	0.5	0.90832

傾向スコアを用いた共変量の調整における バイアスと標準誤差のふるまいについて

松井優作
東京理科大学 大学院 理学研究科 数理情報科学専攻

下川朝有、川崎洋平、宮岡悦良

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

医療研究において、治療群と対照群のような二群における効果の違いを推定するとき、研究デザインは大きく介入研究と観察研究に分けられる。前向きの介入研究では処置を無作為に選ぶことが可能なため、群間において共変量が偏らないことが期待できる。一方、後ろ向きの観察研究では無作為に選ぶことができないため、共変量の偏りを調整する必要がある。共変量の調整法としては、重回帰モデルのあてはめ、層別化、マッチングなどの手法が用いられている。しかしこれらの手法は共変量の数が増えると調整が困難になるなどの問題がある。そこで共変量を傾向スコアと呼ばれる値に集約し調整に用いる手法が提案された。傾向スコアは観察研究において生じる交絡因子の影響を除去し、疑似ランダム化されているような状況を作るために用いられる。本研究では応答変数を二値と仮定し、傾向スコアによる調整法を用いて推定したオッズ比と従来の重回帰モデルのあてはめにより推定したオッズ比をSASによるシミュレーションを用いて比較した。結果として、特定の状況下において傾向スコアによる調整法ではバイアスが大きくなる場合があることが確認された。

SASシステム

動画による統計表現

～新しい統計の要約～

関根　暁史
株式会社 ACRONET／生物統計部

Dynamic statistical graphs

Satoshi Sekine
ACRONET Corp./Biostatistics Dept. Data Science Division

要旨

SAS には動画を簡単に作成できる機能がある。数式に依存しない形で、多くの人に統計の概念を伝えることができるであろう動画の試作品を作成したので紹介する。

キーワード：SAS グラフ，GIF

1. はじめに

SAS には、複数枚の SAS グラフをコマ送りにして動画としてしまう機能が備わっている[1)2)]。この機能を活かし、ほとんど数学を使わない動画の形で統計を表現してみたら、統計初心者にも統計学の根本が伝わると考えた。本論文では「動的な三次元図」「動的な分割表」「動的な・・・」という章立てとして、統計学にも様々な分野があるが、分野にとらわれることなく、それぞれの章に最適と思われるテーマを用意し、そのテーマの説明補助となり得るような動画の見本を試作した。本論文中の動画は、当日発表用スライド（パワーポイント）をご参照頂きたい。

2. 動画作成プログラム

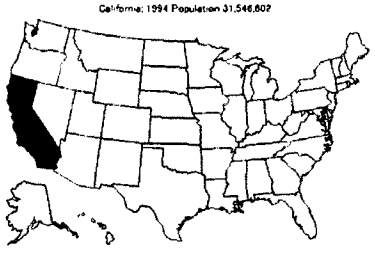


図 0. SAS 社 HP 掲載のプログラム

SAS 社ホームページに図 0 作成のプログラムが掲載されている「http://support.sas.com/kb/25/255.html」。図 0 はアメリカ合衆国の地図における州の色が経時的に塗られていくというものである。同プログラムは拡張子 GIF となる動画を作成して吐き出す。この GIF ファイルはパワーポイントに貼りつければ、スライドショーにすると駆動するので、プレゼンテーションの最中に動画を見せることが可能である。またプレゼンテーションを行う環境は SAS がインストー

ルされている必要もない。同プログラムを小加工して流用するだけで、SAS グラフが描ける人ならば誰でも簡単に動画が作成可能である。よって以降の動画は図 0 の作成プログラムを元に作成することとする。

3. 動的な三次元図

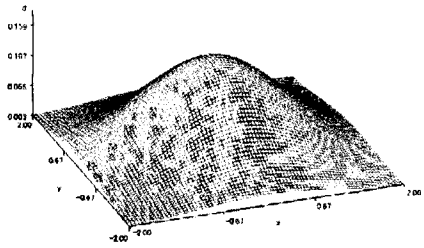


図 1. 二次元正規分布

「SAS によるデータ解析入門」(東京大学出版会)[3]p.135～に掲載のデータ normal は相関係数 r=0.6 の時の 2 次元正規分布を示している。図 0 のプログラムを利用して、この相関係数 r を 0 から 0.9 まで 0.1 ずつ変化させた三次元図を作成することを考える。図 0 のプログラムで回転していたのは&state というマクロ変数のみであったので、相関係数 r=&state.とおいて、state=0 to 0.9 by 0.1 となるデータセット usa を用意して、図 0 のプログラムをそのまま実行すれば図 1 が完成する。本プログラムのソースコードは巻末のプログラム 1 に掲載した。

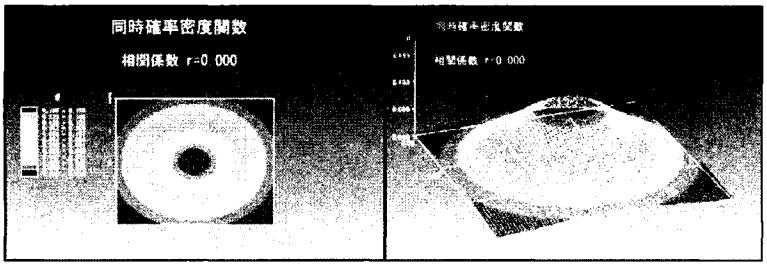


図 2. 二次元正規分布（高品位）

さらに参考文献 4)を参照頂くことで、図 2 のように高品位にすることも可能である。図 2 では 2 枚の別々の図が同時に動いているが、コマ送りの速度を同じにしているので同期して動いて見える。コマ送りの速度は、delay=の値を SAS 側から設定すればよい。パワーポイントにおいて、同時に複数枚の動画を動かいたい際は、ディレイタイムを同じとすることで、無理に 1 枚絵に仕立てる必要はない。本章では確率密度と相関との関係を示す動画として紹介した。

4. 動的な分割表

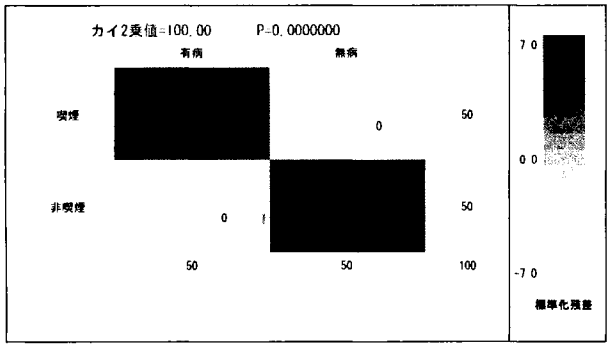


図 3. カイ 2 乗検定

度数とともに色の濃淡が変化する分割表を考えた。(喫煙・非喫煙) × (有病・無病) の 2 ×2 分表（人工データ）であるが、周辺分布を固定しながら各セルを 1 例ずつ変化させていく。セル色の濃淡は標準化残差の値と紐付いている。すなわち色の濃いセルには度数が集中しているし、色の薄いセルは期待値と比較して度数の少ないことが判る。全てのセルの度数が期待値と変わらない場合は、一様の平面が出来る。この一様性が崩れるほどカイ 2 乗値が跳ね上がる。P 値は 5%有意の時、赤字で表現されるが、喫煙・無病が多い方向性で 5%有意になる際は、緑色となる。

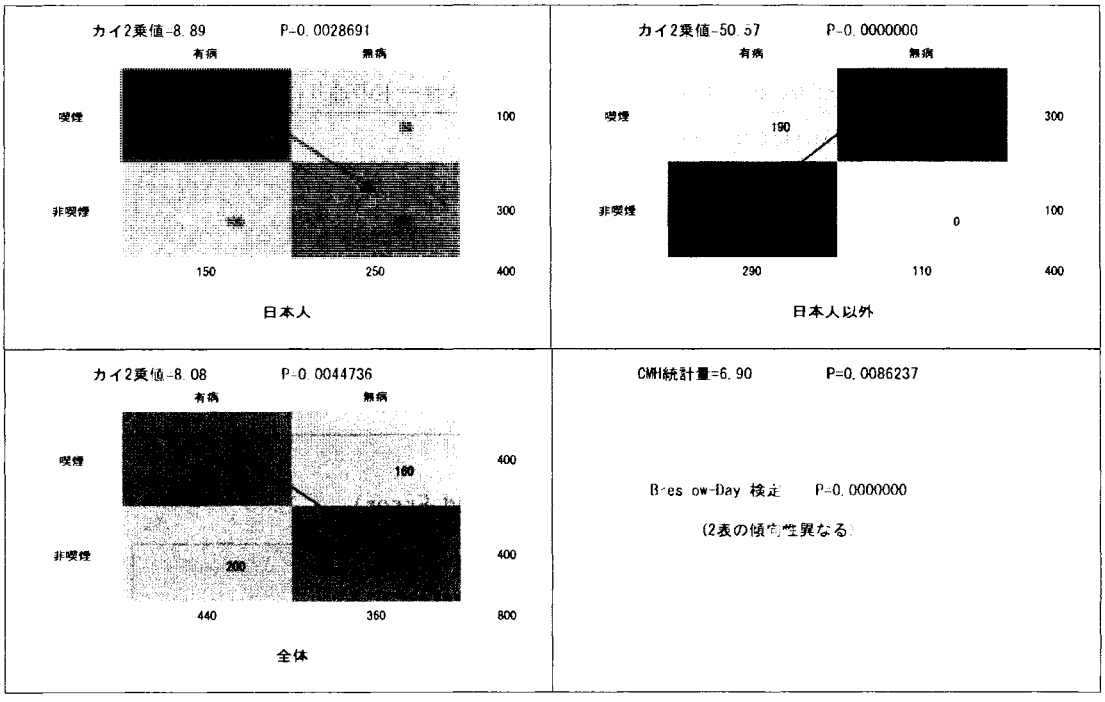
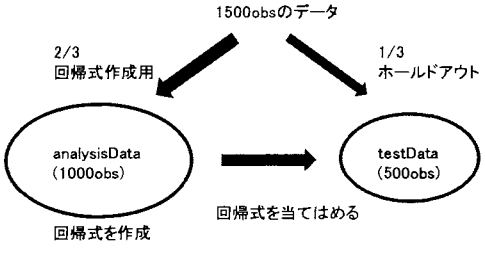


図 4. 層別カイ 2 乗検定

図 3 の概念の分割表を 2 枚同時に動かすことを考える。日本人と日本人以外のそれぞれ 400 例ずつの分割表のセル度数は変化させるが、全体（日本人+日本人以外）800 例のセル度数は全く変化させないようにする。この図 4 の状態で、全体と日本人では 5%有意であるが、日本人以外は喫煙・無病の方向性で 5%有意の状態である。表の傾向性はセル色の濃い部分をたどれば見えてきて、この図中には傾向性の矢印を書き込んでいる。図 4 は、日本人と日本人以外は別の傾向性を持っているにもかかわらず、全体の P 値が有意となっている。これは右下の Breslow-Day 検定にも反映されていて、Breslow-Day 検定が 5%有意の時“(2 表の傾向性異なる)”と赤字で表示するようにした。すなわち当該図表はシンプソンのパラドクスを示したものである。日本人と日本人以外が、最後まで全体と同じ傾向のまま有意にならないという人工的データを作成して本動画を作成した。つまり全体の有意をもたらしているのは、喫煙か非喫煙がではなく、国別という原因が作用しているのではないかという例を示した。

5. 動的な折れ線



1,500obs のデータを analysisData（3 分の 2）と testData（3 分の 1）に分割し、analysis 側で線形重回帰式を作成してその回帰式を test 側に適用することを考える。変数選択を伴う回帰分析において、analysis 側、test 側とともに ASE（残差平方和を N 数で割ったもの）を計算して逐次お互いの ASE を

比較する。analysis 側は最小 2 乗法によって ASE を減らしていくが、test 側は言わば受身的に ASE を計算させられることになる。本データは SAS ヘルプの GLMSELECT の章にあるものを用いているが(Example 42.2 Using Validation and Cross Validation)、本回帰分析は下記のソースコードの通りに行った。

```
proc glmselect data=analysisData testdata=testData;
    class c1 c2 c3(order=data);
    model y =   c1|c2|c3|x1|x2|x3|x4|x5|x5|x6|x7|x8|x9|x10
                |x11|x12|x13|x14|x15|x16|x17|x18|x19|x20 @2
              / selection=stepwise(select = sl)
                hierarchy=single;
run;
```

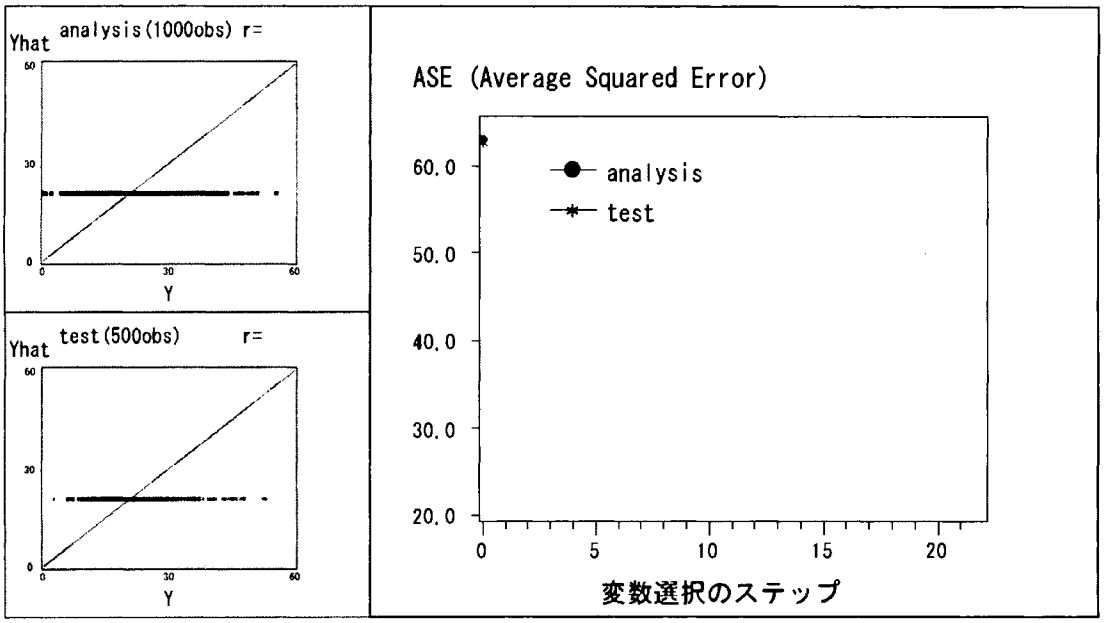


図 5. 線形重回帰分析における過学習の概念

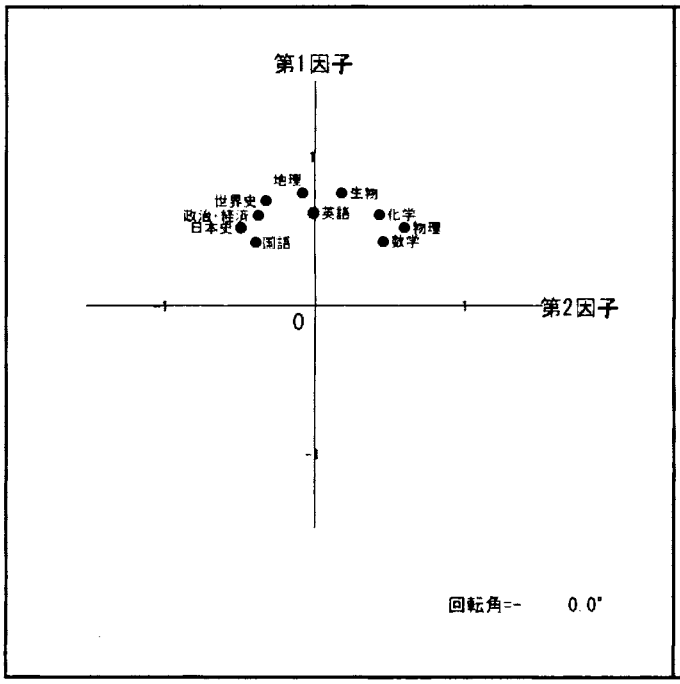
上記プログラムを実行した時の ASE の変動の過程を見たものが図 5 である。左に補助的に Y と Yhat の散布図を相関係数とともに付けている。変数選択のステップが進むごとに analysis 側、test 側ともに ASE は下がり、散布図の分布は 45° の対角線に近づいていく（相関係数は上昇していく）。しかし 10 ステップ目で test 側は ASE が最小(相関係数は 0.790)になった後、11 ステップ以降 ASE は上昇していくことになる。よって analysis 側は過学習をしていることが考えられ、analysis 側のステップは 10 ステップ目付近で止めておくことがバイアス減少のために相応しいと思われる。本動画は回帰分析における過学習の概念を伝えるものであり、11 ステップ目以降になると“Over Learning”と赤字で表示するようにしている。

6. 動的な座標軸

10 教科 50 人分の人工データを用意して因子分析を行って見る。下記教科データは「SAS によるデータ解析入門」[3]p.193 掲載の認知課題データを小加工したものである。

英語	数学	国語	物理	化学	生物	日本史	世界史	地理	政治・経済
63	57	74	56	25	63	66	68	88	78
66	40	83	56	65	70	62	72	76	60
...									
60	52	78	52	80	63	70	54	76	52

本データを nfact=2 の FACTOR プロシジャに供する。プロシジャのデフォルトのまま主成分解を解くものとする。第 1 因子を縦軸、第 2 因子を横軸として因子負荷量の散布図を描く(図 6)。既に因子軸の回転前から第 1 因子は 10 教科の総合得点を、第 2 因子はいわゆる文理を意味していることが想像できる。回転前因子負荷量の分散(2 乗和)は、(第 1 因子，第 2 因子)=(3.722，1.395)であった。この因子軸をバリマックス法によって直交回転させてみる。



回転前因子負荷量

因子負荷量	第1因子	第2因子
英語	0.622	-0.005
数学	0.440	0.456
国語	0.428	-0.387
物理	0.527	0.596
化学	0.617	0.429
生物	0.763	0.180
日本史	0.521	-0.478
世界史	0.705	-0.316
地理	0.762	-0.075
政治・経済	0.604	-0.363
分散(2乗和)	3.722	1.395

回転後因子負荷量

因子負荷量	第1因子	第2因子
英語	0.622	-0.005
数学	0.440	0.456
国語	0.428	-0.387
物理	0.527	0.596
化学	0.617	0.429
生物	0.763	0.180
日本史	0.521	-0.478
世界史	0.705	-0.316
地理	0.762	-0.075
政治・経済	0.604	-0.363
分散(2乗和)	3.722	1.395

図 6. 因子分析における直交回転の概念

動画では-45° 方向に座標軸を回すことによってバリマックス回転をイメージした。バリマックス回転後、(第 1 因子，第 2 因子)=(2.687，2.429)となり分散がより平均化した。「数学」「物理」などに着目すると、回転後因子負荷量の第 1 因子にはほとんど寄与しなくなった。逆に「国語」「日本史」などに着目して見ると第 2 因子には寄与しなくなって、単純構造が得られていることが判る。第 1 因子は“文系能力”を、第 2 因子は“理系能力”を表していると解釈できるので回転の終わりに赤字で表示した。本動画は、回転前・回転後で座標軸の直交性が崩れていないということと、各教科間の因子負荷量の内積(相関性)に変化が無いということを示すために作成した。

7. 動的なデンドログラム

6章の人工データをそのまま用いて、変数のクラスター分析を行うこととする。デフォルトのVARCLUSプロシジャ[5]によりクラスター数を上昇させていく実験を行う。初期状態の分割クラスター数が1の時の分散説明率0.372とは、6章の第1因子の分散3.722（すなわち主成分分析の第1固有値）を教科数10で割った値と一致する。分割クラスター数を2とした時、文系的な第1クラスター(英・国・日・世・地・政)と理系的な第2クラスター(数・物・化・生)に割り付けられる。因子分析のオーソブリク回転によって第1次割付がなされる。回転後因子負荷量の絶対値が第1因子の方により寄与していた教科は第1クラスターに割り付けられ、第2因子の方により寄与していた教科は第2クラスターに割り付けられたのである。実際のアルゴリズムはk-meansクラスタリングによく類似していて、因子分析のオーソブリク回転以降に、主成分分析による第2次割付へと反復されるのであるが、本データではmaxirer=1で全て十分収束してしまうので反復の詳しい解説は割愛する。分割クラスター数=2において、第1クラスターの中での主成分分析の第1固有値は2.739、第2クラスターの中での主成分分析の第1固有値は2.153であるので、合計値4.892を10で割ったものが分散説明率となっており、クラスター数1→2の上昇で、説明率0.372→0.489に上昇した。

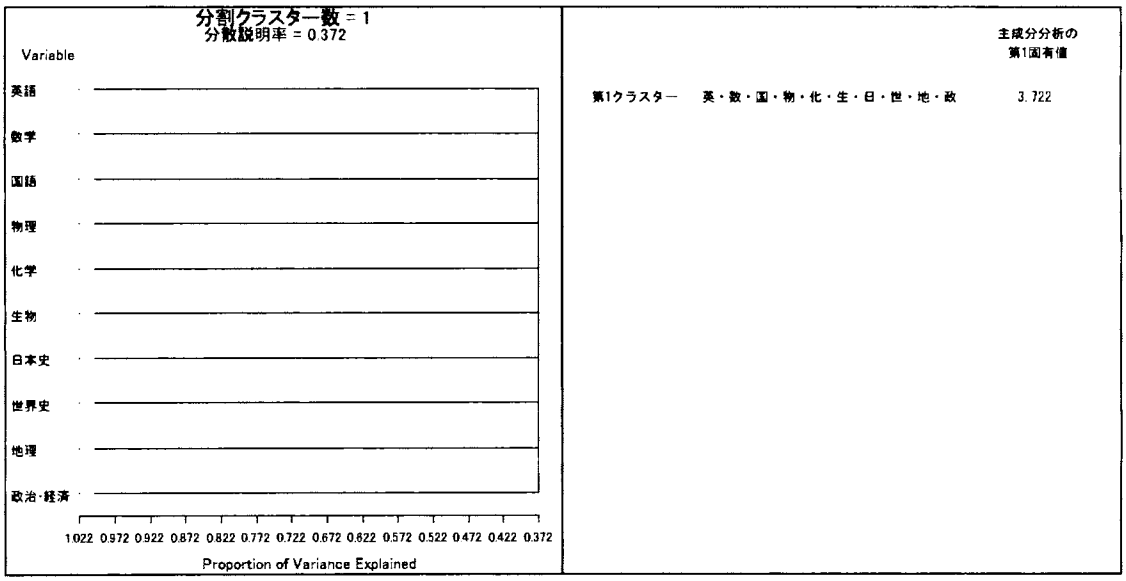


図7. 変数の階層的クラスター分析

その後動画を見ていくと、理系の中でも「数物クラスター」・「生化学クラスター」に割れたり、文系の中でも「言語クラスター(英・国)」が現れたりする。10教科しかないので10個のクラスターまで分割して、分散説明率が元の1となって終了である（固有値もそれぞれ1ずつとなって終わる)。本動画は、VARCLUSプロシジャがクラスター分析とは言っても主成分・因子分析に近い考え方をしていること紹介するために作成して見た。

8. 動的なしきい

ある臨床検査薬を考える。500 例の有病群は正規分布 $N(70,15^2)$に従っており、9,500 例の無病群は正規分布 $N(40,15^2)$に従っている(すなわち有病率 5%)。しきい値を超えた場合を陽性(+)とみなし、それ以外は陰性(-)である。しきい値を 30 から上昇させたとき、感度と特異度の変化を図 8 に表した。しきい値は ROC 曲線上も動いている。

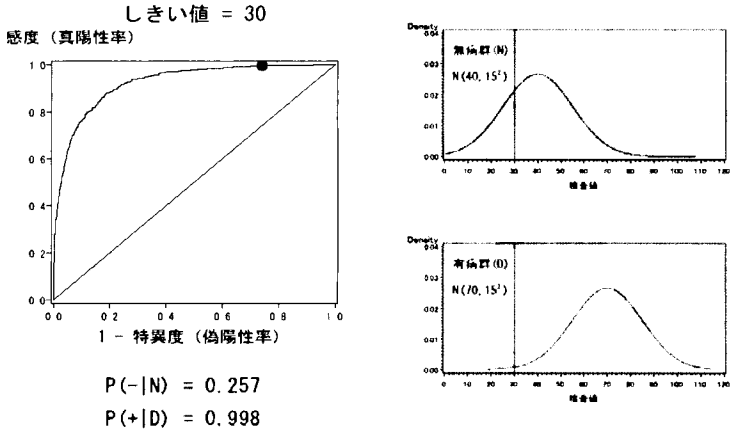


図 8. ROC 曲線としきい値

9. 動的な ROC

500 例の有病群は正規分布に従っており、標準偏差は 15 のまま平均が 55 から 80 に変化させる。9,500 例の無病群は正規分布に従っており、標準偏差は 15 のまま平均が 55 から 30 に変化させる(2 群は等分散としている)。群間差が開くに従って ROC 曲線の AUC が 0.5 から上昇する様子を図 9 に示した。t-検定にも考え方が近いので群間差の t-統計量の表示も添えた。

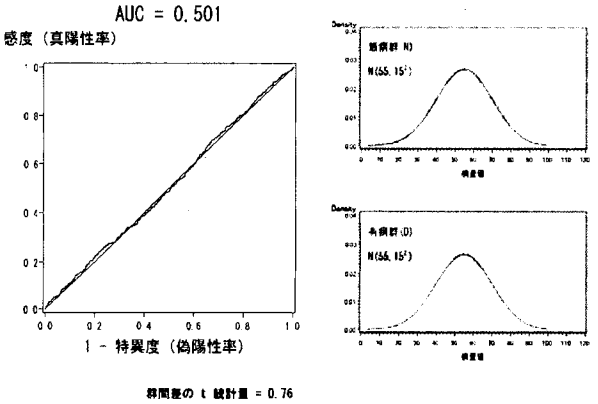
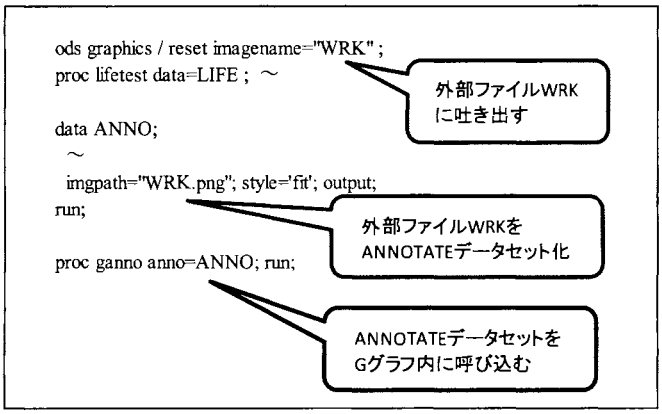


図 9. AUC と群間差の関係

10. 動的なカプランマイヤー図

本章ではカプランマイヤー図を動かすということだけではなく、SG グラフを動画にするということを同時に試みている。goption である GIFANIM Device Driver は残念ながら SG グラフ(すなわち ods graphics on にて出力されるグラフ)はサポートしていない。しかし一工夫することで SG グラフを動画にすることができる[6]。本プログラムのソースコードは巻末のプログラム 2 に掲載した(グラフ中の検定統計量表示部分は、紙面の関係上割愛させて頂いた)が、SG グラフを動画化する手順を以下に記す。



手順 1：ods graphics 機能によって SG グラフを外部ファイル(拡張子 png)として吐き出す

手順 2：外部ファイルをそのまま ANNOTATE データセット化してしまう

手順 3：ANNOTATE データセットを ganno プロシジャに呼び込んで回転させ、あたかも SG プロシジャが回転していることにしてしまう

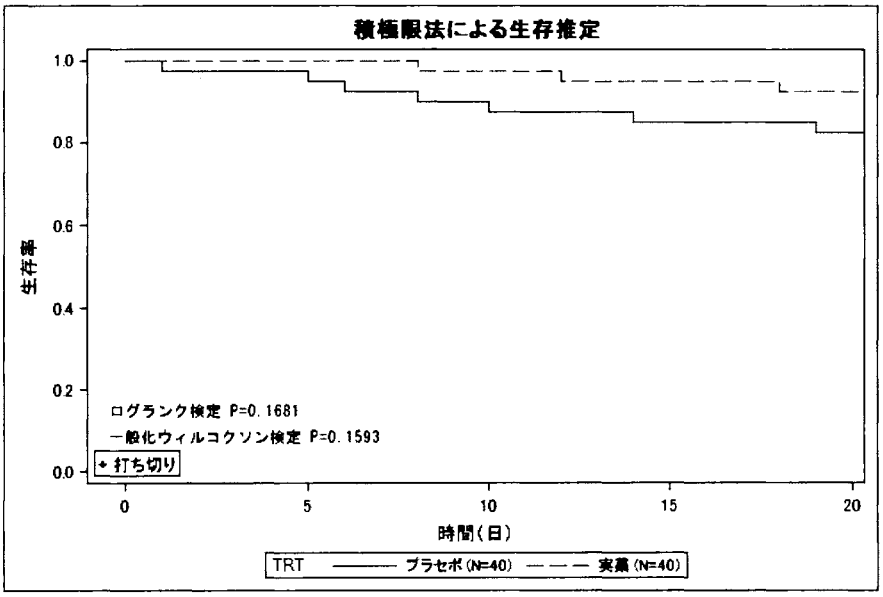


図 10. ログランク検定とウィルコクソン検定の比較

図 10 は、プラセボ群と実薬群の生存率を 20～600 日まで追跡した時の、ログランク検定と一般化ウィルコクソン検定の違いを示した動画である。最初は実薬群の生存率が勝っているように見えるが、300 日付近で生存率が逆転するという人工データを作った。一般化ウィルコクソン検定は比較的初期の差を見ているのに対し、ログランク検定は時間軸の後方の差を見ているのがお判りいただけるとと思う[7]。

11. 動的なクラスター

有名なフィッシャーのアヤメのデータ(1936)を用いて、k-means クラスタリングの概念を動画にすることとした(図 11)。データには 4 変数あるが、二次元の散布図で表現したいため、そのうち花弁の幅・花弁の長さのみを用いることとする。本 150 件のデータを FASTCLUS プロシジャによって 3 分割する。反復数を少なくしたいため、初期シードはそれぞれのクラスターに近いデータを代表値とした。左にはオリジナルの種(セトサ・バーシカラー・バージニカ)ごとの散布図を参考までに置いた。

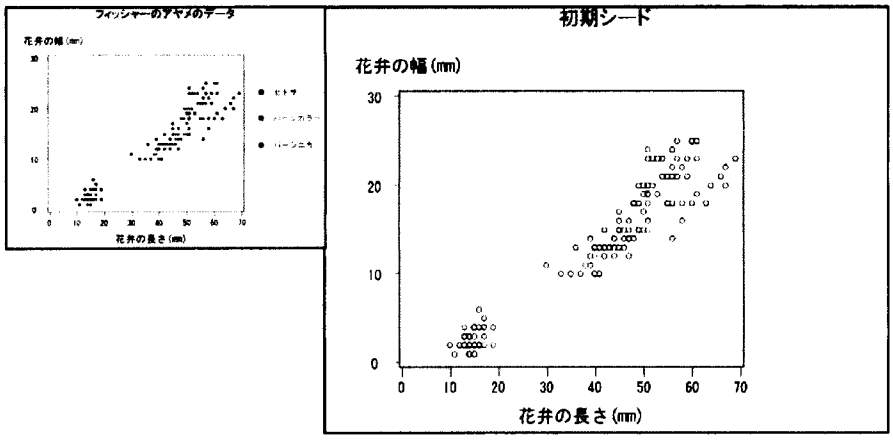


図 11. k-means クラスタリングの概念

動画では先ず初期シード（代表値）が設定され、それぞれ 3 つのシードからユークリッド距離が近い点がそれぞれ 3 つのクラスターに割り付けられる。それぞれのクラスターの重心が計算され、その重心を第 2 のシードとしてクラスタリングが繰り返される。前回の割付と矛盾がなかった場合、収束したとみなし、クラスタリングは終了する。この 1 回の実験は計 4 回の反復で終了し、150 件のうち 8 件が左図と比較して誤答であったが最終クラスターはオリジナルのデータに近い分類となった。

12. まとめ

動画は大量の情報を 1 枚に集約して表現することができる。プレゼンテーションにおいて結果だけでなくプロセスを説明するのに適している。複数枚のタイプの異なる動画を同期的に動かすことで、統計量の多面的な連動を表現することができる。動画とすることで何かシミュレーション（実験的なこと）を行っていることを伝えることができ、説明を簡略化できる。上記の動画は全て SAS9.2 を用いて作成した(※)。SAS で動画を作成することは特殊な外部オプションを必要とせず容易なことなので、ご自身のニーズに合った動画作成にチャレンジして頂きたい。

(※)SAS9.3 以上で実行される場合は、メイン SAS ウィンドウの上部にあるメニューからツール→オプション→プリファレンスを選択、結果タブを表示し、「HTML を作成する」のチェックを外し、「リストを作成する」にチェックを入れて実行すると上手く行くでしょう。

参考文献

1) 長谷川 要 (2002). スピログラフを再現しよう －GIFANIM Device Driver を用いたアニメーション図形の作成－, 日本 SAS ユーザー会

2) 岸本 容司 (2003). SAS グラフによる動く万華鏡の作成, 日本 SAS ユーザー会

3) 竹内 啓 (1994). SAS によるデータ解析入門[第 2 版], 東京大学出版会

4) 関根 暁史 (2012). 色を自在に操る (HSV カラーコードのすすめ), SAS ユーザー総会

5) 岸本 淳司 (1996). 変数のクラスタリング－PROC VARCLUS 再発見－, 日本 SAS ユーザー会

6) Xin Zhang (2013). Extended SAS GIFANIM Device Usage on Table Reporting and Template-Based Graphics, SAS Global Forum

7) 大橋 靖雄 (1995). 生存時間解析 SAS による生物統計, 東京大学出版会

```
/* プログラム 1 */
%macro onestate( state, ds );
data normal;
  r=&state.; pai=3.141593;
  c=(1/(2*pai*(1-r**2)**0.5));
  do x=-2 to 2 by 0.05;
      do y=-2 to 2 by 0.05;
          d=c*exp(-(0.5/(1-r**2)*(x**2-2*r*x*y+y**2)));
          output;
      end;
   end;
run;
proc g3d data=normal;
   plot y*x=d/rotate=20 tilt=40;
run; quit;
%mend;

data usa;
  do state=0 to 0.9 by 0.1; output; end;
run;

data _null_;
   set usa end=done;
   file '～URL 指定～¥normal.sas';
   if _n_ = 1
   then put "filename animmap '～URL 指定～¥NORMAL.gif';" /
             "goptions reset=goptions device=gifanim gsfmode=replace gsfname=animmap xpixels=600
ypixels=400"
             "cback=white iteration=0 delay=150 disposal=background noborder htitle=13pt;";
   else if _n_ = 2
   then put "goptions gsfmode=append;";
   if done then put "goptions gepilog='3B'x;";
   put '%onestate(' state ', usa );';
run;

%inc '～URL 指定～¥normal.sas';
```

```
/* プログラム 2 */
proc format;
   value trt 1="実薬 (N=40)" 2="プラセボ (N=40)";
run;

data LIFE;
   input TRT TIME CENSOR @@;
   format TRT trt.;
   label TRT="治療" TIME="時間（日）" CENSOR="打ち切り";
   cards;
   1 8    0     2 1    0
   1 12   0     2 5    0
   1 18   0     2 6    0
   1 24   0     2 8    0
   1 36   0     2 10   0
   1 48   0     2 14   0
   1 68   0     2 19   0
   1 84   0     2 22   0
   1 95   0     2 32   0
   1 102 0      2 34   0
   1 109 0      2 40   0
   1 118 0      2 46   0
   1 132 0      2 50   0
   1 144 0      2 54   0
   1 156 0      2 62   0
   1 168 0      2 64   0
   1 174 0      2 66   0
   1 184 0      2 68   0
   1 192 0      2 72   0
   1 198 0      2 74   0
   1 219 0      2 80   1
   1 220 1      2 82   0
   1 232 0      2 86   0
   1 244 0      2 96   0
   1 252 0      2 105 0
   1 264 0      2 120 0
   1 270 0      2 160 0
   1 290 0      2 280 0
```

```
   1 300 0      2 300 0
   1 320 0      2 401 0
   1 333 0      2 501 0
   1 350 0      2 600 0
   1 364 0      2 610 0
   1 400 0      2 650 0
   1 464 0      2 678 0
   1 502 0      2 694 0
   1 555 0      2 700 0
   1 559 0      2 701 0
   1 601 0      2 800 0
   1 602 0      2 900 0
   ;
run;

%macro onestate( state );
ods graphics / reset width=6in height=4in imagename="WRK" ;
proc lifetest data=LIFE maxtime=&state.;
   time TIME*CENSOR(1);
   strata TRT;
run;

data ANNO;
   length function style $ 32 ;
   retain xsys ysys '3' hsys '3' when 'a';
   function='move'; x=0; y=0; output;
   function='image'; x=100; y=100;
   imgpath="WRK.png"; style='fit'; output;
run;

proc ganno anno=ANNO; run;
%mend;

data usa;
   do state=20 to 600 by 20; output; end;
run;
```

```
data _null_;
    set usa end=done;
    file '～URL 指定～¥KM.sas';
    if _n_ = 1
    then put "filename animmap '～URL 指定～¥カプランマイヤー.gif';" /
                "goptions reset=goptions device=gifanim gsfmode=replace gsfname=animmap xpixels=601
ypixels=401"
                "cback=white iteration=0 delay=200 disposal=background border htitle=13pt ;";
    else if _n_ = 2
    then put "goptions gsfmode=append;";
    if done then put "goptions gepilog='3B'x;";
    put '%onestate(' state ');';
run;

%inc '～URL 指定～¥KM.sas';
```

伝統芸能実演家の動的データベースの作成

坂部 裕美子
公益財団法人 統計情報研究開発センター 研究開発本部

Creation of the Dynamic Database about the Traditional-Performing-Arts Performers

Yumiko Sakabe
Research-and-Development Headquarters,
Statistical Information Institute for Consulting and Analysis

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

伝統芸能の実演者に関する長期データ中から、ある任意の年次時点で活動していた実演家のみを抜き出し、比較用の属性を付加して動的データベースを作成するプログラムを開発した。その適用例を紹介する。

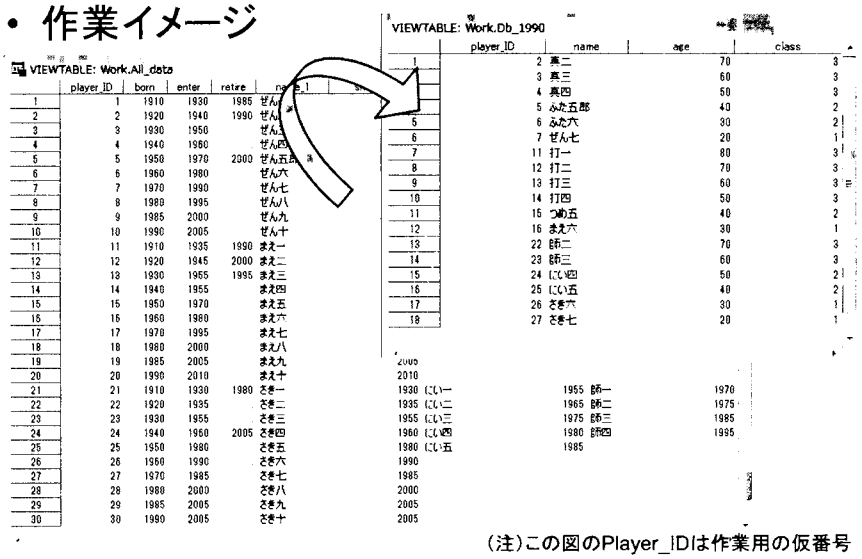
キーワード:arrayステートメント、mergeステートメント、updateステートメント、rankプロシジャ

1プログラムの必要性

- 筆者は、伝統芸能の上演傾向等を年度ごとに集計・比較するという研究をしている
- 年度ごとの公演概況集計結果の参照用資料として、当時の実演家の付帯情報をまとめたDBが必要
- 実演者リストは、現実的には単年度版が市販されることが多く、観客としてはその方が使いやすい。しかし、データ整備・更新作業面からは、積年版の単一ファイルの方が好都合

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

- 作業イメージ



(注)この図のPlayer_IDは作業用の仮番号

2 共通作業の存在

- 実際に「歌舞伎」「落語」「宝塚」の実演家DBを開発していく過程で、すべてに共通する作業過程があり、プログラムを共有できるのではないかと考えた
  - 襲名および改名...該当年度の名前で表示する
  - 昇進...該当年度の階級で表示する
- 「共通プログラム」と「分野ごとの独自指標の算出」を経て、長期DBから特定の年度の実演家DBを抽出表示する

3 共通プログラム

① 全演者データから該当年度に活動していた者のみを抽出し、年齢を計算する

Player_ID...各演者のユニークID
落語＝入門順
宝塚＝成績順

```
%let year=1990;
data db_&year;
  set all_data;
  format player_ID name age class <F1><F2>;
    if retire ne . and &year > retire then delete;
    if enter > &year then delete;
    %determ
    age= &year-born;
    keep player_ID name age class <F1><F2>;  run;
```

② その当時の階級と名称で表示する

```
%macro determ;
    class=0;
  %do i=1 %to 5;
    if name_&i ne '' and shumei_&i=<&year then
      name=name_&i;
    if shoshin_&i=<&year then
      class+1;
  %end;
%mend;
```

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

4 芳野特化プログラム①－歌舞伎

- 共通プログラム中の要修正箇所
  - 階級昇進なし、改名（襲名）あり
- 独自項目（実演者DBからの付加項目）
  - 親は誰か（将来的には「歌舞伎俳優か」も考慮する）
  - 特定の顕彰制度についての受賞歴
- 独自算出指標（DB内で算出するもの）
  - 家の格（受賞歴を使用）

「家の格」を定義

- 「芸術院会員」「人間国宝」「文化功労者」「文化勲章」のうち2つ以上の顕彰歴のある俳優は「名家」と考えたい(=ランク1)
- 上記の条件を満たす俳優の息子は「特別視」したい(=ランク2)
- これらの条件に該当する俳優にフラグを付ける

  → この「家の格」を使用した集計を、過去のSASユーザー会で報告済み

(1) 4賞受賞者を抽出

```
data ds01;
 set db_&year (keep=player_ID prize1-prize4);
 array prz{4} prize1-prize4;
 flg01=0;
  do i=1 to 4;
  flg01+(prz{i} ne . and prz{i} =< &year );
  end;
  if flg01>=2 then output;
run;
```

「該当年次における受賞状況」とすることで、役者自身の「ランク2」から「ランク1」への昇格を反映できる
(例:菊五郎7)

(2) 該当者へフラグ付け

1. 親が該当するケースへmergeする

```
data ds01;
  set ds01(keep=player_ID);
  prz01=2; run;
proc sort data=db_&year;
  by shisho;
data ds02_1;
  merge db_&year ds01
  (rename=(player_ID=shisho));
  by shisho;
run;
```

2. 当人が該当するケースへupdateする

```
data ds01;
  set ds01;
  prz01=1;run;
proc sort data=db_&year;
  by player_ID;
data ds02_2;
  update ds02_1 ds01;
  by player_ID;
run;
```

両方に該当するケースは、
より上位の1に上書きされる

5 分野特化プログラム②ー落語

- 共通プログラム中の要修正箇所
  - なし(階級、名称、年齢すべて処理が必要)
- 独自項目
  - 師匠は誰か
- 独自算出指標
  - 入門以降の経過年数
  - 真打に抜擢昇進したか
  - 弟子／兄弟弟子の人数

抜擢昇進フラグ付け

- 落語家の階級＝前座 / 二つ目 / 真打
- 抜擢昇進：二ツ目の落語家が、先輩を飛び越えて真打に昇進するケース
  - 将来性を高く見込まれて抜擢が行われることが多い。実際に、昇進後の寄席の出番は、そうでないものより有意に多い
  - 「抜いた人数の多寡」よりも「抜擢されたか否か」の方がその後の活躍への影響が大きい

  → 過去のSASユーザー会で分析報告済み

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

1. 未昇進者の真打昇進年に仮の値を入れる

```
data ds01;
  set all_data(keep=player_ID shumei_3);
   if shumei_3=. then shumei_3=2999;
run;
```

入門後当該年までの死亡・廃業者も考慮し、正確な抜擢状況を把握する（今後の研究で使用する可能性あり）

2. 真打昇進順位が入門順位より上の人にフラグを付ける

```
proc rank data=ds01 out=ds02  ties=low;
  var shumei_3;  ranks shoshin;
run;
data ds02;   set ds02;    if shoshin<player_ID then flg01=1;
run;
```

弟子・兄弟弟子について

- 落語家のライフサイクルを考察する上で有意義な指標たりうると考えられる
- 弟子がいる≒(1)ある程度成功している
  (2)後進の育成期に入っている
- 「弟子の数」や「自身が惣領弟子であるか否か」も指標の一つとなり得る？(今後の検証が必要)

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

1. 兄弟弟子の連番をふる

```
proc rank data=db_&year out=ds03_1  ties=low ;
var enter;  ranks deshi_no ; by shisho;
run;
```

2. 自分の弟子の人数を付与する

```
proc freq data=db_&year;   tables shisho /  noprint out= ds04 ;   run;
data ds03_2;  merge ds03_1 ds04
(drop=percent  rename=(shisho=player_ID  count=deshi));
  by player_ID;  run;
```

6 分野特化プログラム③ー宝塚

- 共通プログラム中の要修正箇所
  - 階級昇進なし、年齢なし、改名あり(僅かに存在)
- 独自項目
  - 所属する組、男役/娘役の別
  - トップ就任の有無および就任年
  - 新人公演の主演および本公演以外の主演状況
- 独自算出指標
  - 入団以降の経過年数
  - 本公演以外の公演種別主演回数

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

組替え情報の集計

- 過去の組替え回数は、その後の活動内容と関連があると考えられるので、別途項目として立てておく

この部分の作業がこれ以降で繰り返しになるので、マクロを定義する。

```
data ds01;
 set db_&year (keep=player_ID  c_class_year: );
 array c_class{*} c_class_year: ;
  c_class_t=0;
 do i=1 to dim(c_class);
   c_class_t+(c_class{i} ne . and c_class{i} =< &year );
 end;
run;
```

公演主演状況

- 新人公演・バウホール公演・その他の劇場の主演経験歴を集計する

```
data ds02;
 set db_&year (keep=player_ID  shin_top: vow_top: other_top:  );

  %gsum(shin)
  %gsum(vow)
  %gsum(other)

run;
```

```
%macro gsum(var);
 array &var{*} &var_top_: ;
 &var_main=0;
 do i=1 to dim(&var);
  &var_main+(&var{i} ne . and &var{i} =< &year );
 end;
%mend;
```


まとめ

- 収録データの規模は、「全データ」で最大4500件程度、「抽出データ」で最大400件程度（うち、独自指標の算出が必要なのは数十件）の見込みなので、処理時間の問題なく実用化できると考えられる
- Outputに関しては未開拓だが、もともと結果閲覧が目的なので、将来的に着手できればと考えている
- この分野の研究にSASを使っている人は皆無だと思うので、珍奇な事例報告としてご紹介しました。まだプログラムに改良の余地はあると思います。

SASでCDISC SDTMデータを効率的に利用するためにDefine-XMLのメタデータを活用する

富永 一宏

イーピーエス株式会社 統計解析1部

Use of Define-XML Metadata for Efficiently Making CDISC SDTM Datasets in SAS Programing

Kazuhiro Tominaga
Statistics Analysis Department 1, EPS Corporation

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

SASのXMLエンジンの機能を利用してDefine-XMLからMetadataを読み取り、SDTMの親ドメインとSUPPQUALの結合した後に非標準変数の属性をデータ利用者が想定していたものに復元する方法。

キーワード:SDTM, define.xml, Define-XML, XMLMap

イントロダクション

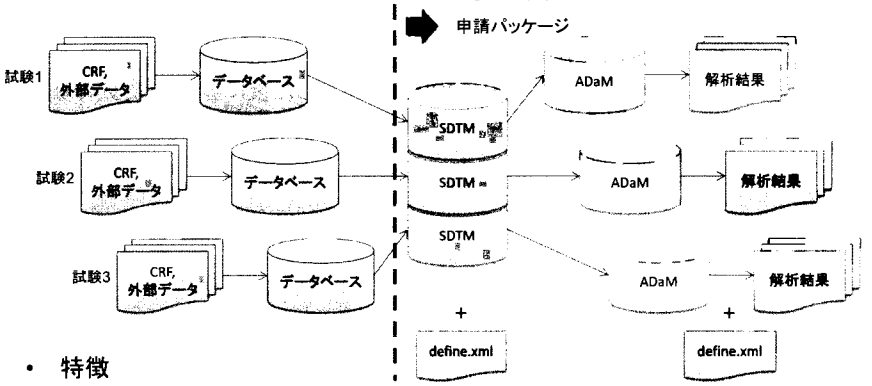
- CDISC標準
  - FDA
    - 2004年からCDISC標準によるデータ提出が試みられてきた。
  - 日本
    - 申請データ提出フォーマットとして採用された。
    - 2016年から提出が本格的に開始される予定。

- CDISC標準に準拠したデータを作成するために
  - 例えば、米国を中心にCDISC標準に準拠したデータの作成方法が検討されてきた。
    - 疾患領域別ガイド、Questionnaireガイド、Controlled Terminologyの充実
  - さらに近年はdefine.xmlを作成する便利なツールも登場。

これまでは、データ作成者側の立場としての試みが注目されてきた

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

SDTMの特徴



- 特徴
  - 入力/管理から分離されたデータ
  - 元々のコンセプトは申請データ提出用にRawデータを高度に標準化したもの。
  - 他方、高度に標準化しすぎた部分(SUPPQUALなど)はデータ利用者には使いづらい。

目標

- 目的

今後CDISC標準に準拠したデータが増えてくるならば、
次はデータ利用者側が「効率的にデータを利用できる方法」が重要になる。

- 課題
  - 試しにSDTMデータの利用者の立場になって扱ってみると・・・
    - define.xmlのメタデータを直接プログラムから読み込みたくなることが分かった。
    - SASでこれを実現できるのか？
- 今回の目標
  - 今回は以下の簡単なシチュエーションを例に検討してみる。

SDTMの親ドメインにSUPPQUALを再結合するときに発生する課題と、
define.xmlを用いた解決方法をSASプログラムで実現する。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

SDTMのSUPPQUALのイメージ

SDTMデータを利用するときの姿

USUBJID	XXSEQ	XXTESTCD	XXORRES	XXSTRESC	XXVAR1 (Label 1)	XXVAR2 (Label 2)
ABC123	1	XXXX	X.X	X.X	X	Y

非標準変数

SDTMデータとして格納した姿

XX

USUBJID	XXSEQ	XXTESTCD	XXORRES	XXSTRESC
ABC123	1	XXXX	X.X	X.X

+

SUPPXX

USUBJID	IDVAR	IDVARVAL	QNAM	QLABEL	QVAL
ABC123	XXSEQ	1	XXVAR1	Label 1	X
ABC123	XXSEQ	1	XXVAR2	Label 2	Y

USUBJID	XXSEQ	XXVAR1 (Label 1)	XXVAR2 (Label 2)
ABC123	1	X	Y

非標準変数を正規化し、
SDTM標準変数のみにする
(縦積みに転置する)

SDTMデータ利用者はこの手順を
逆にたどる必要がある

SUPPQUALから転置した時の課題

SUPPXX

USUBJID	IDVAR	IDVARVAL	QNAM	QLABEL	QVAL
ABC123	XXSEQ	1	XXVAR1	Label 1	X
ABC123	XXSEQ	1	XXVAR2	Label 2	Y

```
PROC TRANSPOSE DATA=SUPPXX PREFIX=_;
 VAR QVAL;
 BY STUDYID RDOMAIN USUBJID IDVARVAL;
 ID QNAM;
 IDLABEL QLABEL;
 WHERE IDVAR='XXSEQ';
RUN;
```

QVALの型(文字)と長さが
転置後の変数に継承される。

元々の非標準変数の
型、長さ、順序が復元できない。

XX

USUBJID	XXSEQ	...	...	...
ABC123	1	...	...	...

+

USUBJID	XXSEQ	XXVAR1	XXVAR2
ABC123	1	X	Y

define.xmlからメタデータを
読み取る必要

非標準変数はSUPPQUALで値として存在するので
欲しい情報はdefine.xmlの
Value Level Metadataとして格納されている

Define-XML

- 特徴
  - CDISC標準の一つで、メタデータの格納方法を規定したもの
    - XMLファイルで実装する。
    - 現在、バージョンはV1.0とV2.0の2つが存在する。
  - 簡単に言うと
    - ブラウザーで閲覧するデータ定義書。

Demographics (DM) dm.xpt

Variable	Label	Key	Type	Length	Code List / Controlled Terms	Origin	Role	Source/Derivation/Comments
STUDYID	Study Identifier		text	12		CRF Page	IDENTIFIER	
DOMAIN	Domain Abbreviation		text	2		Assigned	IDENTIFIER	
USUBJID	Unique Subject Identifier		text	11		Derived	IDENTIFIER	Concatenation of STUDYID, DM.SITEID and DM.SUBJID
SUBJID	Subject Identifier for the Study		text	4		CRF Page	TOPIC	
RFSTDTC	Subject Reference Start Date/Time		date	10	ISO8601	Derived	RECORD QUALIFIER	Date/time of first study drug treatment derived from EX
RFENDTC	Subject Reference End Date/Time		date	10	ISO8601	Derived	RECORD QUALIFIER	Date/time of last study drug treatment derived from EX

Define-XML

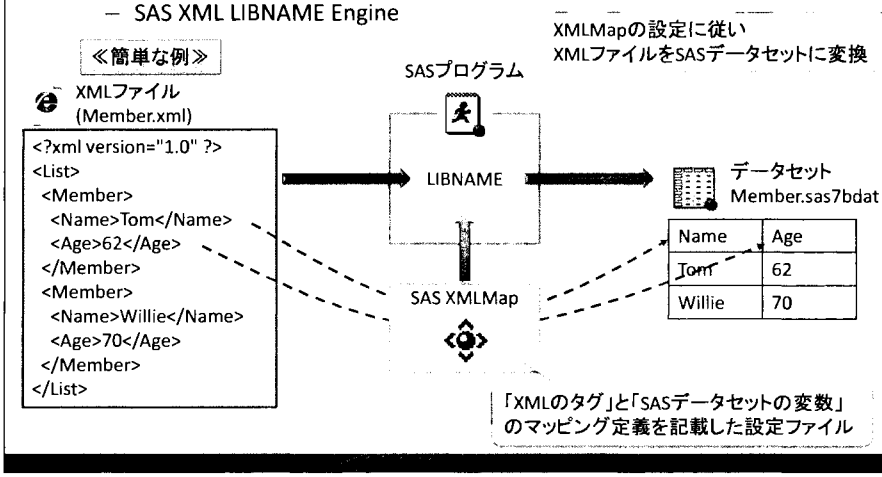
- XMLファイルのメリット・デメリット
  - メリット
    - プログラムから読み取り可能(Machine Readable)
      - コンピュータで表示、閲覧できるという意味ではない。
      - SASなどのプログラミングによって情報を読み取れるという意味。
  - デメリット
    - 作成するためには、IT技術に詳しいプログラマが必要。
    - ただし、最近はdefine.xml作成ツールが公開されてきて、敷居が下がってきた。
      - 書籍「Implementing CDISC Using SAS - An End-to-End Guide」のSASマクロ
        http://support.sas.com/publishing/authors/shostak.html
      - 富士通「tsClinical Define.xml Generator」
        http://jp.fujitsu.com/solutions/life/cdisctool/

V1.0.0ではValue Level Metadata に対するName, SASFieldNameの 実装方法にまだ問題がある

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

SAS XML LIBNAME Engine

- SASからXMLファイルを読み取る方法
  - SAS XML LIBNAME Engine



SAS XML LIBNAME Engine

- 特徴
  - SAS XMLMap
    - XMLファイルのエレメントをSASデータセットの変数にマッピングするための設定ファイル
    - 設定ファイルの仕様を理解する必要がある。
      - SAS 9.4 XML LIBNAME Engine: User's Guide
        http://support.sas.com/documentation/cdl/en/engxml/64990/HTML/default/viewer.htm#titlepage.htm
    - 自動生成ツール
      - 「SAS XML Mapper」というソフトを使うとXMLMapファイルの雛形を簡単に作成できる。
        http://support.sas.com/downloads/browse.htm?fil=&cat=12
  - XMLファイルの書き込み機能が弱い
    - 単純な書き込み機能しかないため、define.xmlのような複雑なXMLファイルを作成する事はできない。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

XMLMapファイル

```
<?xml version="1.0" encoding="UTF-8"?>
<SXLEMAP name="AUTO_GEN" version="2.1">
 <NAMESPACES count="0"/>
 <TABLE description="Member" name="Member">
  <TABLE-PATH syntax="XPath">/List/Member</TABLE-PATH>

  <COLUMN name="Name">
   <PATH syntax="XPath">/List/Member/Name</PATH>
   <TYPE>character</TYPE>
   <DATATYPE>string</DATATYPE>
   <LENGTH>10</LENGTH>
  </COLUMN>

  <COLUMN name="Age">
   <PATH syntax="XPath">/List/Member/Age</PATH>
   <TYPE>numeric</TYPE>
   <DATATYPE>integer</DATATYPE>
  </COLUMN>
 </TABLE>
</SXLEMAP>
```

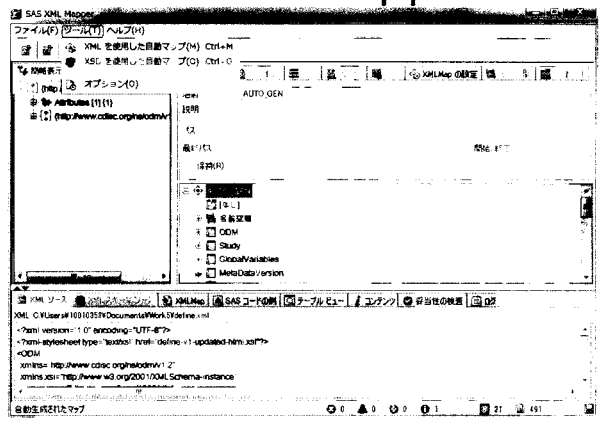
XMLのエレメント指定は
XPathで記述する

データセットの定義
ここで指定したTABLE-PATH
を基準にOBSが発生する

Nameのマッピング定義

Ageのマッピング定義

SAS XML Mapper



- define.xmlを読み込むとエラーする場合の対処法
  - ODMタグのxsi:schemaLocation属性を削除すると改善する場合がある。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

LIBNAMEの設定

- 設定方法

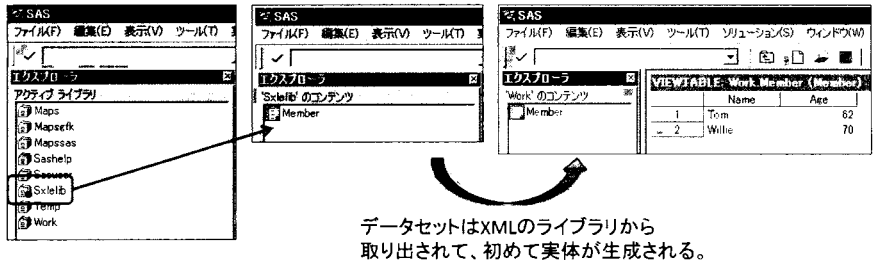
```
filename sxlelib "C:¥Temp¥Member.xml";
filename sxlemap "C:¥Temp¥Member.map";
libname sxlelib xmlv2 xmlmap = sxlemap;

data member;
 set sxlelib.member;
run;
```

読み込むXMLファイルの指定

XMLMapファイルの指定

LIBNAMEの設定



データセットはXMLのライブラリから
取り出されて、初めて実体が生成される。

非標準変数の属性を復元する

- SUPPQUALを転置した時の課題
  - 非標準変数の以下の情報が復元できない
    - 型、長さ、順序
- define.xmlを読み込む
  - 非標準変数の属性情報はValue Level Metadataにある
    - SAS XML LIBNAME Engineでこの情報を読み取る
    - 読み取った情報を利用して、非標準変数の属性を復元する。
- 以下ではDefine-XML V1.0を想定

Define-XML V2.0を読み込む場合の注意事項
- XMLMapファイル
  namespaceの値と変数ラベルの仕様が異なる。
- SASプログラム
  QNAMではなく、QVALからValueListのIDを抽出する。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

define.xmlで読み込む箇所

- 必要な情報

SDTMデータを利用するときの全変数の情報一覧

Datasets

Dataset	Description	Class
DM	Demographics	Special Purpose
SUPPDM	Supplemental Qualifiers for DM	Relationship
...		

DM

Variable	Label	Type	Length
STUDYID	Study Identifier	text	12
DOMAIN	Domain Abbreviation	text	2
USUBJID	Unique Subject Identifier	text	11
SUBJID	Subject Identifier for the Study	text	4
...			

SUPPDM

Variable	Label	Type	Length
STUDYID	Study Identifier	text	12
RDOMAIN	Related Domain Abbreviation	text	2
USUBJID	Unique Subject Identifier	text	11
IDVAR	Identifying Variable	text	8
IDVARVAL	Identifying Variable Value	text	200
QNAM	Qualifier Variable Name	text	8
QLABEL	Qualifier Variable Label	text	40
QVAL	Data Value	text	200
...			

+

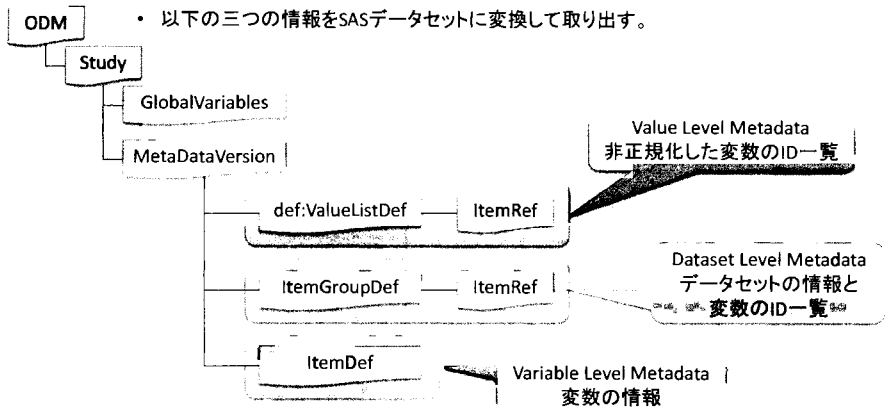
SUPPDMのValue Level Metadata

Source Variable	Value	Label	Type	Length
QNAM	SAFETY	Safety Population Flag	text	1
QNAM	ITT	Intent to Treat Population Flag	text	1
...	...			

最終的に欲しい情報

define.xmlで読み込む箇所

- XMLの構造
  - 以下は、SUPPQUALを転置した後に、属性を再設定するために必要となる部分。
  - 以下の三つの情報をSASデータセットに変換して取り出す。



医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

define.xml 用のXMLMapファイル

XMLMapファイルの骨子

```
<?xml version="1.0" encoding="UTF-8"?>
<SXLEMAP version="2.1">
 <NAMESPACES count="3">
  <NS id="1" prefix="">http://www.cdisc.org/ns/odm/v1.2</NS>
  <NS id="2" prefix="def">http://www.cdisc.org/ns/def/v1.0</NS>
  <NS id="3" prefix="xlink">http://www.w3.org/1999/xlink</NS>
 </NAMESPACES>

 <TABLE name="Value">...</TABLE>

 <TABLE name="Dataset">...</TABLE>

 <TABLE name="Variable">...</TABLE>
</SXLEMAP>
```

Namespace設定

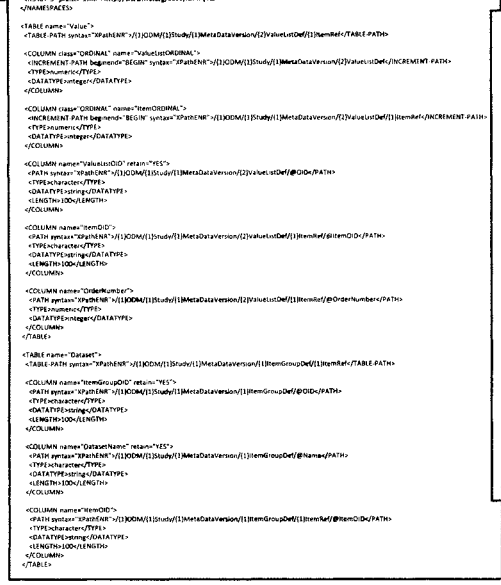
Value Level Metadata
非正規化した変数のID一覧

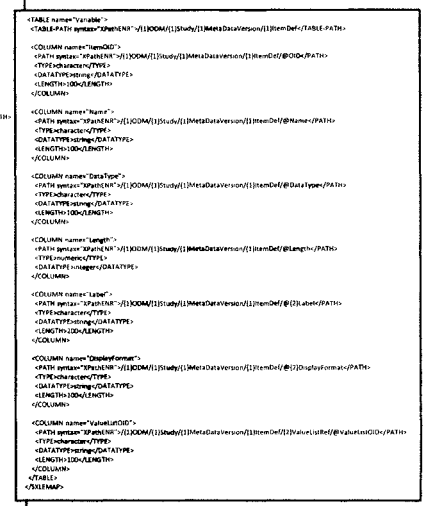
Dataset Level Metadata
データセットの情報と
変数のID一覧

Variable Level Metadata
変数の情報

Define-XML V1.0用のXMLMapファイル

ValueMetaDefine1.map

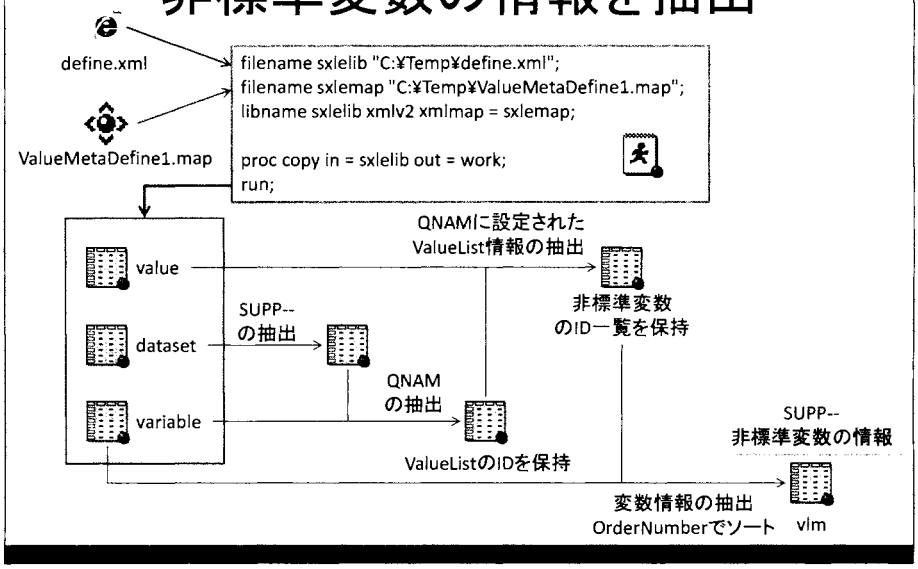


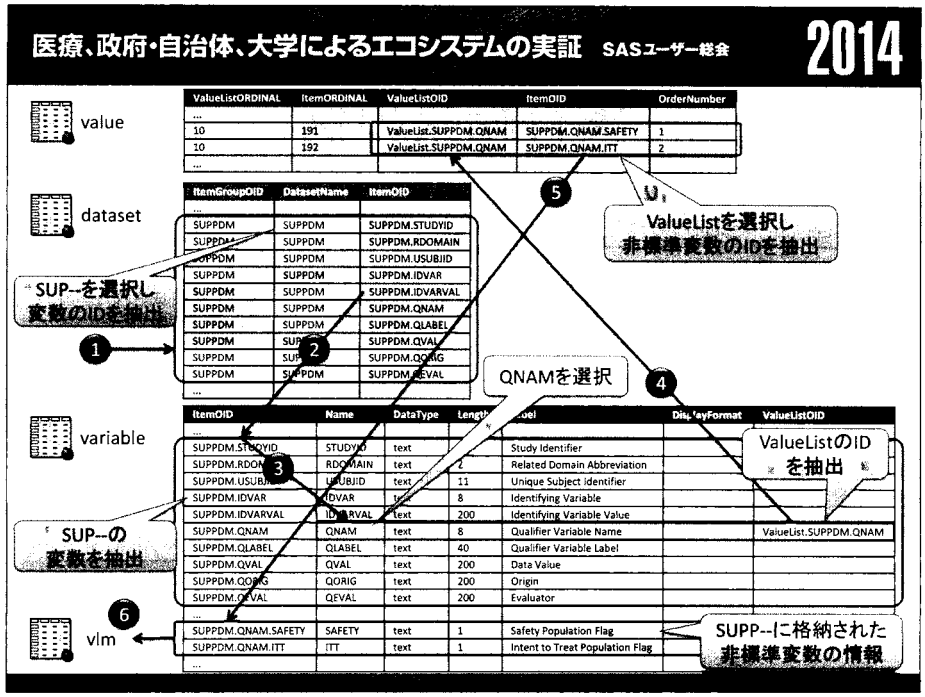


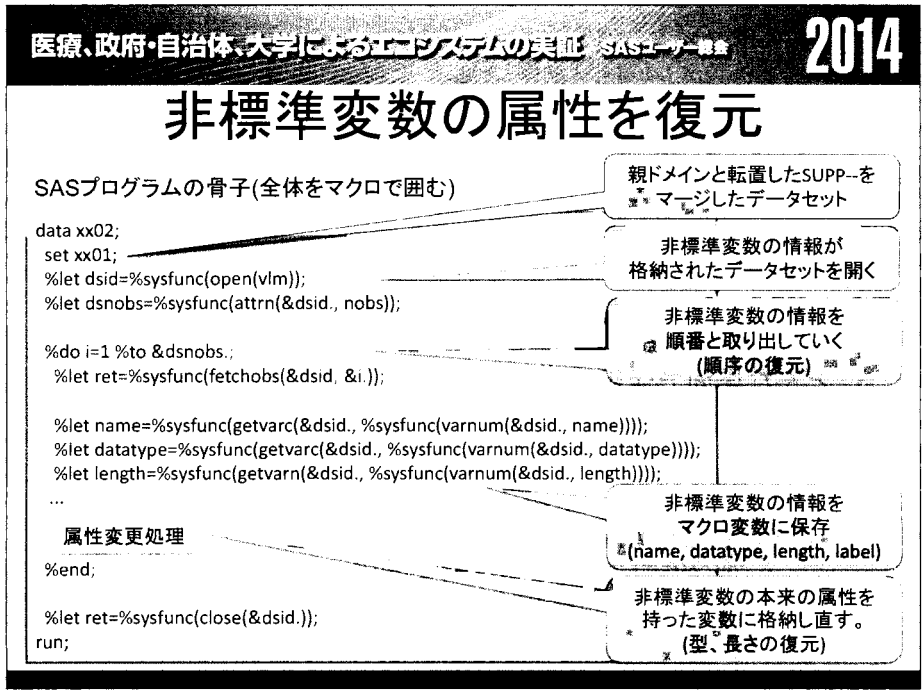
SAS XML Mapperで雛形を作成してから
細部を整えると簡単に作れる。

医療、政府・自治体、大学によるエコシステムの実証　SASユーザー総会　2014

非標準変数の情報を抽出




医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会
2014
value
dataset
variable
vlm
SUP--を選択し
変数のIDを抽出
SUP--の
変数を抽出
QNAMを選択
ValueListを選択し
非標準変数のIDを抽出
ValueListのID
を抽出
SUPP--に格納された
非標準変数の情報
1
2
3
4
5
6


医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会
2014
非標準変数の属性を復元
SASプログラムの骨子(全体をマクロで囲む)
data xx02;
set xx01;
%let dsid=%sysfunc(open(vlm));
%let dsnobs=%sysfunc(attrn(&dsid., nobs));
%do i=1 %to &dsnobs.;
%let ret=%sysfunc(fetchobs(&dsid, &i.));
%let name=%sysfunc(getvarc(&dsid., %sysfunc(varnum(&dsid., name))));
%let datatype=%sysfunc(getvarc(&dsid., %sysfunc(varnum(&dsid., datatype))));
%let length=%sysfunc(getvarn(&dsid., %sysfunc(varnum(&dsid., length))));
...
属性変更処理
%end;
%let ret=%sysfunc(close(&dsid.));
run;
親ドメインと転置したSUPP--を
マージしたデータセット
非標準変数の情報が
格納されたデータセットを開く
非標準変数の情報を
順番と取り出していく
(順序の復元)
非標準変数の情報を
マクロ変数に保存
(name, datatype, length, label)
非標準変数の本来の属性を
持った変数に格納し直す。
(型、長さの復元)

実行結果の例

- DM+SUPPDMの例

Value	Label	Type	Length
...			
_ITT	Intent to Treat Population Flag	text	200
_SAFETY	Safety Population Flag	text	200
...			

転置して生成された非標準変数部分

この例では以下が復元している
- 長さ
- 順序

Value	Label	Type	Length
...			
SAFETY	Safety Population Flag	text	1
ITT	Intent to Treat Population Flag	text	1
...			

以下の一連のプログラムをマクロ化すれば自動化できる。
- SUPPQUALを転置し、親ドメインと結合
- define.xmlの読み込み
- 非標準変数の属性を復元

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

まとめ

- 今回の例から分かったこと
  - SDTM利用者の視点で考えると
    - SDTMのデータセットのみでは、利用時のすべての変数情報が揃っていない。
      - Value Level Metadataに関してはdefine.xmlにしかない状態。
    - define.xmlをSASプログラムで読み込める事は、SDTMを効率的に利用するための足掛かりとなる。
- どんな場面で活用できるのか
  - SDTMの親ドメインにSUPPQUALを再結合するときの問題(今回の例)
    - SDTMデータから直接レビューする人は、毎回のように再結合する必要があるため、自動化できると便利。
    - ADaMを作成する人は、特にSUPPQUALに多数の非標準変数があるときに便利。
  - Dataset-XML V1.0の読み込み、書き込み
    - CDISCが作成している次世代のデータセット交換用ファイル形式。
    - データ利用者はDataset-XMLをSASデータセットに変換する必要がある。
    - Dataset-XMLはDefine-XMLが必須の仕様になっているため、相互変換プログラムにはdefine.xmlの読み込みが必須。(実際にSASプログラムを作って試してみました)

ODS markupを使ったADaM define-xmlの作成

坂上　拓　　矢嶋　友也　　西本　優美
株式会社 中外臨床研究センター バイトメトリクス部　データサイエンスグループ

Creation of ADaM define-xml using SAS ODS markup

Taku Sakaue　Tomoya Yajima　Yumi Nishimoto
Biometrics Dept. Data Science Group, Chugai Clinical Research Center., LTD

1. 要旨

承認申請時の CDISC 標準に準拠した臨床電子データの提出義務化を控え，これら臨床電子データ定義書としての位置づけとなる define-xml の作成は，解析業務プロセスの一つに組み込まれることが予想される．

ADaM や解析帳票を作成する解析プログラミング業務のプロセスを考えた場合，define-xml を作成するためのデータソースとして，プログラム開発者向けのプログラム仕様書を用いると，ADaM や解析帳票のプログラム仕様と define-xml 間の整合性を保持する面で大きなメリットがある．しかしながら，define-xml の要素や属性を考慮すると，プログラム仕様書には解析とは直接関係のない多大な情報を記入する必要があり，これらを埋めるための時間を割くことは，解析を主として行う担当者にはストレスとなる．

本発表は，当社で検討中の Microsoft Excel で作成されたプログラム仕様書をデータソースとした，SAS ODS markup を使った define-xml 作成方法と，define-xml の要素や属性に関する記入箇所を極力減らしたプログラム仕様書を紹介する．

キーワード： ADaM define-xml, プログラム仕様書, SAS ODS markup

2. ADaM define-xml 作成までの処理概要

当社での ADaM define-xml 作成までの工程を以下に示す。

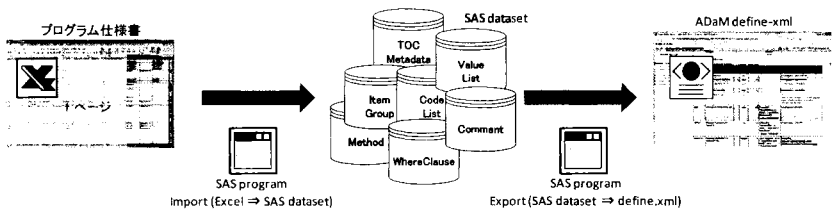


図 I. ADaM define-xml 作成工程

3.1 プログラム仕様書

プログラム仕様書は Excel を使って作成しており，ADaM や解析帳票のプログラム仕様と共に，define-xml を作成するために必要な情報も記載される．プログラム仕様書の構成を表 1 に示す．"Description"列には Excel への記載ルールと共に，define-xml へ展開するための仕様を記載しており，図 1 の中で示されているプログラム仕様書から SAS dataset へインポートするプログラムと，SAS dataset から define-xml へエクスポートするプログラムのプログラム仕様に相当する内容となる．

表 1. プログラム仕様書の構成

Sheet Name [Description]	Row / Column Name	Description
STUDY [試験情報]	FILE ID	試験単位の File ID を設定 (試験で一貫した ID を使用) 1). ODM (FileOID) 2). [FILE ID]
	STUDY ID	試験単位の Study ID を設定 (試験で一貫した ID を使用) 1). Study (OID) 2). [STUDY ID]
	STUDY NAME	試験名 (試験で一貫した試験名を使用) 1). GlobalVariables/StudyName 2). [STUDY NAME]
	PROTOCOL NAME	プロトコル名 (試験で一貫したプロトコル名を使用) 1). GlobalVariables/ProtocolName 2). [PROTOCOL NAME]
	STUDY DESCRIPTION	試験内容 (試験で一貫した記述) 1). GlobalVariables/StudyDescription 2). [STUDY DESCRIPTION]
SUPPORTDOC [参照文書]	ID	文書単位の ID 1). def:Leaf (ID) 2). LF.[ID]
	SUPPDOC	Supplemental Document として定義するかどうか (Y or null) "Y"が指定されている文書は，Supplemental Document として定義 1). def:SupplementalDoc/def:DocumentRef (LeafID) 2). LF.[ID]
	TITLE	文書名 1). def:leaf/def:title 2). [TITLE]
	PATH	文書への相対パス 1). def:leaf (xlink:href) 2). [PATH]
TOC_METADATA [ADaM の一般情報 : ADaM Metadata]	NAME	ADaM dataset 名 1-1). ItemGroupDef (OID) 2-1). IG.[NAME] 1-2). ItemGroupDef (SASDatasetName) 2-2). [NAME] 1-3). ItemGroupDef (def:CommentOID) 2-3). COM.[NAME] 1-4). ItemGroupDef (def: ArchiveLocationID) 2-4). LF.[NAME] 1-5). ItemGroupDef/def:leaf/def:title 2-5). [NAME].xpt
	REPEATING	被験者辺りに複数レコードが発生するデータかどうか (Yes or No) 1). ItemGroupDef (REPEATING) 2). [REPEATING]
	DESCRIPTION	Dataset の説明 1). ItemGroupDef/Description/TranslatedText 2). [DESCRIPTION]
	STRUCTURE	Dataset の構造特性 1). ItemGroupDef (def: Structure) 2). [STRUCTURE]
	CLASS	データ分類 ItemGroupDef (def: Class)
	PATH	データセット(xpt ファイル)の相対パス 1). ItemGroupDef/def:leaf (xlink:href) 2). [PATH]
	DOCUMENTATION	データセットに関連する補足説明 1). def:CommentDef/Description/TranslatedText 2). [DUCMENTATION]
	REFDOC	参照文書 1 記載方法 (デリミタ：/) : 文書単位の ID / 参照方法 (PhysicalRef / NamedDestination) / 参照先 a) a) 参照方法が PhysicalRef の場合は，ページ番号 or ページ範囲 (範囲の場合は##－## ハイフンで区切る)．NamedDestination の場合は，ファイル参照名 文書単位の ID :

Sheet Name [Description]	Row / Column Name	Description 1). 定義情報が記述する define-xml Element (Attribute) 2). define-xml 設定情報
		1). def:CommentDef/def:DocumentRef (leafID) 2). LF.[文書単位の ID] 参照方法： 1). def:CommentDef/def:DocumentRef/def:PDFPageRef (Type) 2). [参照方法] 参照先： 1-1). def:CommentDef/def:DocumentRef/def:PDFPageRef (PageRefs) 2-1). [参照先] 1-2). def:CommentDef/def:DocumentRef/def:PDFPageRef (First Page) 2-2). [範囲指定 (## - ##)をハイフンで分割した 1 番目の要素] 1-3). def:CommentDef/def:DocumentRef/def:PDFPageRef (Last Page) 2-3). [範囲指定 (## - ##)をハイフンで分割した 2 番目の要素]
	REFDOC2	参照文書 2 REFODC と同様
	REFDOC3	参照文書 3 REFODC と同様
ADaM_[Dataset Name] (e.g. ADaM_ADSL, ADaM_ADAE) [ADaM 仕様 : ADaM Variable Metadata]	VARIABLE	変数名 1-1). ItemGroupDef/ItemRef (ItemOID) 2-1). ID."Dataset Name".[VARIABLE] 1-2). ItemGroupDef/ ItemRef (MethodOID) 2-2). MT."Dataset Name". [VARIABLE] 1-3). ItemDef (OID) 2-3). IT."Dataset Name".[VARIABLE] 1-4). ItemDef (Name) 2-4). [VARIABLE] 1-5). ItemDef (SASFieldName) 2-5). [VARIABLE] * "Dataset Name"は，シート名から取得
	KEY	Key 情報 (1, 2, 3 ... n) 1). ItemGroupDef/ItemRef (KeySequence) 2). [KEY]の順番に従い，[VARIABLE]をカンマ区切りで結合したもの
	LIST	Value Level Metadata で定義対象かどうか (Y or null) "Y"が指定されている文書は，Value Level Metadata に定義 1). ItemDef/def:ValueListRef (ValueListOID) 2). LV."Dataset Name".[VARIABLE] * "Dataset Name"は，シート名から取得
	REQ	変数の Core 情報 (Yes or No) 変数の Core 情報が"Req"の場合は"Yes"を設定 1). ItemGroupDef/ItemRef (Mandatory) 2). [REQ]
	LABEL	変数ラベル 1). ItemDef/Description/TranslatedText 2). [LABEL]
	TYPE	変数型 (text or integer or float or date or datetime) 1). ItemDef (DataType) 2). [TYPE]
	LENGTH	変数長 1). ItemDef (Length) 2). [LENGTH]
	DIGIT	有効桁数 1). ItemDef (SignificantDigits) 2). [DIGIT]
	DISPFMT	表示形式 (SAS format 名) 1). ItemDef (def:DisplayFormat) 2). [DISPFMT]
	CNTRLTERM	Terminology 名 1). ItemDef/CodeListRef (CodeListOID) 2). CL.[CNTRLTERM]
	DERIVTYPE	データ由来 (CRF or Derived or Assigned or Protocol or eDT or Predecessor) 1). ItemDef/def:Origin (Type) 2). [DERIVTYPE]
	METHTYPE	導出タイプ (Computation or Imputation) DERIVETYPE 列で"Derived"が指定されている場合に設定 1). MethodDef (Type) 2). [METHTYPE]
	SOURCEVAR	データソースとなる変数名 DERIVETYPE 列で" Predecessor"が指定されている場合に設定 1). ItemDef/def:Origin/Description/TranslatedText 2). [SOURCEVAR]
	DERIVATION	導出ロジック 1). MethodDef/Description/TranslatedText 2). [DERIVATION]
	DERIVATION (JPN)	導出ロジック（日本語） 日本語での補足説明用．define-xml には行かない情報
	REFDOC	参照文書 1 記載方法： 文書単位の ID / 参照方法 (PhysicalRef / NamedDestination) / 参照先 a) a) 参照方法が PhysicalRef の場合は，ページ番号 or ページ範囲 (範囲の場合は## – ## ハイフンで区切る)．NamedDestination の場合は，ファイル参照名 文書単位の ID： 1). MethodDef/def:DocumentRef (leafID) 2). LF.[文書単位の ID]

		参照方法： 1). MethodDef/def:DocumentRef/def:PDFPageRef (Type) 2). [参照方法] 参照先： 1-1). def:CommentDef/def:DocumentRef/def:PDFPageRef (PageRefs) 2-1). [参照先] 1-2). def:CommentDef/def:DocumentRef/def:PDFPageRef (First Page) 2-2). [範囲指定 (## - ##)をハイフンで分割した 1 番目の要素] 1-3). def:CommentDef/def:DocumentRef/def:PDFPageRef (Last Page) 2-3). [範囲指定 (## - ##)をハイフンで分割した 2 番目の要素]
	REFDOC2	REFDOC と同様
	REFDOC3	REFDOC と同様
CODELIST [Control Terminology]	CNTRLTERM	Terminology 名 (ADaM_[Dataset Name].CNTRLTERM と同じ値) 1). CodeList (OID) 2). CL.[CNTRLTERM]
	DESCRIPTION	Terminology の説明 1). CodeList (Name) 2). [DESCRIPTION]
	CLALIAS	コードリストのコード名 (NCI-EVS Control Terminology Code) 1). CodeList/Alias (Name) 2). [CLALIAS]
	TYPE	Terminology の各コードのデータ型 (text or float or integer) 1). CodeList (DataType) 2). [TYPE]
	CODEDVALUE	コード化された値 1-1). CodeList/EnumeratedItem (CodedValue) 2-1). [CODEDVALUE] 1-2). CodeList/CodeListItem (CodedValue)* 2-2). [CODEDVALUE] * DECODEVALUE 列に値が設定されている場合のみ
	DECODEVALUE	コード化される前の値 1). CodeList/CodeListItem/Decode/TranslatedText 2). [DECODEVALUE]
	RANK	コードの順位 1-1). CodeList/EnumeratedItem (RANK) 2-1). [RANK] 1-2). CodeList/CodeListItem (RANK)* 2-2). [RANK] * DECODEVALUE 列に値が設定されている場合のみ
	CALIAS	コード化された値のコード名 1-1). CodeList/EnumeratedItem/Alias (Name) 2-1). [CALIAS] 1-2). CodeList/CodeListItem/Alias (Name)* 2-2). [CALIAS] * DECODEVALUE 列に値が設定されている場合のみ
	CONTEXT	コードの由来 (nci:ExtCodeID) NCI-EVS 由来のコードの場合は"nci:ExtCodeID"を設定 1-1). CodeList/EnumeratedItem/Alias (Context) 2-1). [CONTEXT] 1-2). CodeList/CodeListItem/Alias (Context)* 2-2). [CONTEXT] * DECODEVALUE 列に値が設定されている場合のみ
VALUELIST [Value Level Metadata]	DATASET	設定値の条件が分岐する変数を有するデータセット (ADaM_[Dataset Name]シートの LIST 列で"Y"が設定されている変数を有するデータセット) 1-1). def:ValueListDef (OID) 2-1). VL.[DATASET].*[VARIABLE]* 1-2). def:ValueListDef/ItemRef (ItemOID) 2-2). VL.[DATASET].*[VARIABLE].[NAME]* 1-3). def:ValueListDef/ItemRef (MethodOID) 2-3). MT.[DATASET].*[VARIABLE].[NAME]* 1-4). def:ValueListDef/ItemRef/def:WhereClauseRef (WhereClauseOID) 2-4). WC.[DATASET].*[VARIABLE].[NAME]* 1-5). def:WhereClauseDef (OID) 2-5). WC.[DATASET].*[VARIABLE].[NAME]* 1-6). def:WhereClauseDef/RangeCheck (def:ItemOID) 2-6). IT.[DATASET].*[CHECKVAR]* * 斜体は同シートの別列の値
	VRIABLE	設定値の条件が分岐する変数 (ADaM_[Dataset Name]シートの LIST 列で"Y"が設定されている変数) 1-1). def:ValueListDef (OID) 2-1). VL.*[DATASET]*.[VARIABLE] 1-2). def:ValueListDef/ItemRef (ItemOID) 2-2). VL.*[DATASET]*.[VARIABLE].*[NAME]* 1-3). def:ValueListDef/ItemRef (MethodOID) 2-3). MT.*[DATASET]*.[VARIABLE].*[NAME]* 1-4). def:ValueListDef/ItemRef/def:WhereClauseRef (WhereClauseOID) 2-4). WC.*[DATASET]*.[VARIABLE].*[NAME]* 1-5). def:WhereClauseDef (OID) 2-5). IT.*[DATASET]*.[VARIABLE].*[NAME]* * 斜体は同シートの別列の値
	NAME	Value Level Metadata 内で定義情報が一意となる名称

Sheet Name [Description]	Row / Column Name	Description 1). 定義情報が記述される define-xml Element (Attribute) 2). define-xml 記述情報
		1-1). def:ValueListDef/ItemRef (ItemOID) 2-1). VL.*[DATASET].[VARIABLE]*.[NAME] 1-2). def:ValueListDef/ItemRef (MethodOID) 2-2). MT.*[DATASET].[VARIABLE]*.[NAME] 1-3). def:ValueListDef/ItemRef/def:WhereClauseRef (WhereClauseOID) 2-3). WC.*[DATASET].[VARIABLE]*.[NAME] 1-4). def:WhereClauseDef (OID) 2-4). IT.*[DATASET].[VARIABLE]*.[NAME] * 斜体は同シートの別列の値
	CHECKVAR	条件分岐の評価元となる値が格納された変数 1). def:WhereClauseDef/RangeCheck (def:ItemOID) 2). IT.*[DATASET]*.[CHECKVAR] * 斜体は同シートの別列の値 * 同一セル内に複数の値が存在する場合，該当変数分の def:WhereClauseDef/RangeCheck を発生させる
	COMPARATOR	条件分岐の比較演算子 (LT or LE or GT or GE or EQ or NE or IN or NOTIN) 1). def:WhereClauseDef/RangeCheck (Comparator) 2). [COMPARATOR] * 同一セル内に複数の値が存在する場合，該当変数分の def:WhereClauseDef/RangeCheck を発生させる
	CHECKVALUE	条件分岐の比較値 1). def:WhereClauseDef/RangeCheck/CheckValue 2). [CHECKVALUE] * 同一セル内に複数の値が存在する場合，該当変数分の def:WhereClauseDef/RangeCheck を発生させる
	DERIVATION	条件に係る補足説明 1). MethodDef/Description/TranslatedText 2). [DERIVATION]
	REFDOC	参照文書 1 記載方法： 文書単位の ID / 参照方法 (PhysicalRef / NamedDestination) / 参照先 a) a) 参照方法が PhysicalRef の場合は，ページ番号 or ページ範囲 (範囲の場合は## – ## ハイフンで区切る)．NamedDestination の場合は，ファイル参照名 文書単位の ID： 1). MethodDef/def:DocumentRef (leafID) 2). LF.[文書単位の ID] 参照方法： 1). MethodDef/def:DocumentRef/def:PDFPageRef (Type) 2). [参照方法] 参照先： 1-1). def:CommentDef/def:DocumentRef/def:PDFPageRef (PageRefs) 2-1). [参照先] 1-2). def:CommentDef/def:DocumentRef/def:PDFPageRef (First Page) 2-2). [範囲指定 (## - ##)をハイフンで分割した 1 番目の要素] 1-3). def:CommentDef/def:DocumentRef/def:PDFPageRef (Last Page) 2-3). [範囲指定 (## - ##)をハイフンで分割した 2 番目の要素]
	REFDOC2	REFDOC と同様
	REFDOC3	REFDOC と同様
SELECTIONCRITERIA [Analysis Result Metadata "Selection Criteria"の条件定義]	DISPLAYID	データの抽出条件の定義されている帳票 ID (ARM_[Display Name]で定義している[DISPLAY IDENTIFIER]と同じ値) 1-1). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/rm:AnalysisDatasets/arm:AnalysisDataset/def:WhereClauseRef (WhereClauseOID) 2-1). WC.[DISPLAYID].*[DATASET]* 1-2). def:WhereClauseDef (OID) 2-2). WC.[DISPLAYID].*[DATASET]* * 斜体は同シートの別列の値
	DATASET	抽出対象データ ([CHECKVAR]が格納されているデータセット名) 1-1). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/rm:AnalysisDatasets/arm:AnalysisDataset/def:WhereClauseRef (WhereClauseOID) 2-1). WC.*[DISPLAYID]*.[DATASET] 1-2). def:WhereClauseDef (OID) 2-2). WC.*[DISPLAYID]*.[DATASET] * 斜体は同シートの別列の値
	CHECKVAR	条件分岐の評価元となる値が格納された変数 1). def:WhereClauseDef/RangeCheck (def:ItemOID) 2). IT.[DATASET].*[CHECKVAR]* * 斜体は同シートの別列の値 * 同一セル内に複数の値が存在する場合，該当変数分の def:WhereClauseDef/RangeCheck を発生させる
	COMPARATOR	条件分岐の比較演算子 (LT or LE or GT or GE or EQ or NE or IN or NOTIN) 1). def:WhereClauseDef/RangeCheck (Comparator) 2). [COMPARATOR] * 同一セル内に複数の値が存在する場合，該当変数分の def:WhereClauseDef/RangeCheck を発生させる
	CHECKVALUE	条件分岐の比較値 1). def:WhereClauseDef/RangeCheck/CheckValue 2). [CHECKVALUE] * 同一セル内に複数の値が存在する場合，該当変数分の def:WhereClauseDef/RangeCheck を発生させる

ARM_[Display Name] (e.g. ARM_TABLE_14-3.01) [解析帳票の仕様：Analysis Result Metadata] *シート名の[Display Name]は arm:AnalysisResultDisplays/arm:ResultDisplay (OID)に設定	DISPLAY IDENTIFIER	帳票 ID 1-1). arm:AnalysisResultDisplays/arm:ResultDisplay (OID) 2-1). RD.[DISPLAY IDENTIFIER] 1-2). arm:AnalysisResultDisplays/arm:ResultDisplay (Name) 2-2). [DISPLAY IDENTIFIER] 1-3). arm:AnalysisResultDisplays/arm:ResultDisplay/def:DocumentRef (leafID) 2-3). LF.[DISPLAY IDENTIFIER] 1-4). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult (OID) 2-4). AR.[DISPLAY IDENTIFIER] 1-5). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:AnalysisDatasets (def:CommentOID) 2-5). COM.[DISPLAY IDENTIFIER] 1-6). def:CommentDef (OID) 2-6). COM.[DISPLAY IDENTIFIER] * 1-5), 1-6)は解析に使用するデータセットの抽出条件等のコメント用
	REFDOCID (DISPLAY IDENTIFIER)	参照文書 1 記載方法： 文書単位の ID / 参照方法 (PhysicalRef / NamedDestination) / 参照先 a) a) 参照方法が PhysicalRef の場合は，ページ番号 or ページ範囲 (範囲の場合は## – ## ハイフンで区切る)．NamedDestination の場合は，ファイル参照名 文書単位の ID： 1). arm:AnalysisResultDisplays/def:DocumentRef (leafID) 2). LF.[文書単位の ID] 参照方法： 1). arm:AnalysisResultDisplays/def:DocumentRef/def:PDFPageRef (Type) 2). [参照方法] 参照先： 1-1). arm:AnalysisResultDisplays/def:DocumentRef/def:PDFPageRef (PageRefs) 2-1). [参照先] 1-2). arm:AnalysisResultDisplays/def:DocumentRef/def:PDFPageRef (First Page) 2-2). [範囲指定 (## - ##)をハイフンで分割した 1 番目の要素] 1-3). arm:AnalysisResultDisplays/def:DocumentRef/def:PDFPageRef (Last Page) 2-3). [範囲指定 (## - ##)をハイフンで分割した 2 番目の要素]
	REFDOCID (DISPLAY IDENTIFIER) 2	REFDOCID (DISPLAY IDENTIFIER)と同様
	REFDOCID (DISPLAY IDENTIFIER) 3	REFDOCID (DISPLAY IDENTIFIER)と同様
	DISPLAY NAME	表示名 (解析名称) 1). arm:AnalysisResultDisplays/arm:ResultDisplay/Description/TranslatedText 2). [DISPLAY NAME]
	RESULT IDENTIFIER	帳票名 1). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/Description/TranslatedText 2). [RESULT IDENTIFIER]
	PARAMETER	解析に使用するパラメータ名 1). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult (ParameterOID) 2). IT.*[DATASET]*.[PARAMETER] * 斜体は同シートの別列の値
	ANALYSIS VARIABLE	解析に使用する変数名 (複数変数が存在する場合はカンマで区切って指定．[DATASET]で指定するデータセットと変数の指定順番の対応をとる) 1). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:AnalysisDatasets/arm:AnalysisDataset/arm:AnalysisVariable (ItemOID) 2). IT.*[DATASET]*.[ANALYSIS VARIABLE] * 斜体は同シートの別列の値 * 同一セル内に複数の変数が指定されている場合は，該当変数分の arm:AnalysisVariable を発生させる
	REASON	解析理由 1). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult (AnalysisReason) 2). [REASON]
	PURPOSE	解析目的 1). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult (AnalysisPurpose) 2). [PURPOSE]
	DATASET	解析に使用するデータセット名 (複数解析に用いる場合はカンマで区切って指定，[ANALYSIS VARIABLE]で指定した変数と指定順番の対応をとる) 1-1). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:AnalysisDatasets/arm:AnalysisDataset (ItemGroupOID) 2-1). IG.[DATASET] 1-2). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:AnalysisDatasets/arm:AnalysisDataset/arm:AnalysisVariable (ItemOID) 2-2). IT.[DATASET].*[ANALYSIS VARIABLE]* 1-3). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:AnalysisDatasets/arm:AnalysisDataset/arm:AnalysisVariable/def:WhereClauseRef (WhereClauseOID) 2-3). WC.*[DISPLAY IDENTIFIER]*.[DATASET] * 斜体は同シートの別列の値 * 同一セル内に複数のデータセットが指定されている場合は，該当データセット分の arm:AnalysisDataset を発生させる
	DATASET COMMENT	解析に使用するデータセットの抽出条件等に関するコメント 1). def:CommentDef/Description/TranslatedText 2). [DATASET COMMENT]
	DOCUMENTATION	解析要件や補足説明 1). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:Documentation/Description/TranslatedText 2). [DOCUMENTATION]

Sheet Name [Description]	Row / Column Name	Description 1). 定義情報が関連する define-xml Element (Attribute) 2). define-xml 設定情報
	REFDOCID (DOCUMENTATION)	解析要件や補足説明の参照文書 1 記載方法： 文書単位の ID / 参照方法 (PhysicalRef / NamedDestination) / 参照先 a) a) 参照方法が PhysicalRef の場合は，ページ番号 or ページ範囲 (範囲の場合は## – ## ハイフンで区切る)．NamedDestination の場合は，ファイル参照名 文書単位の ID： 1). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:Documentation/def:DocumentRef (leafID) 2). LF.[文書単位の ID] 参照方法： 1). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:Documentation/def:DocumentRef/def:PDFPageRef (Type) 2). [参照方法] 参照先： 1-1). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:Documentation/def:DocumentRef/def:PDFPageRef (PageRefs) 2-1). [参照先] 1-2). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:Documentation/def:DocumentRef/def:PDFPageRef (First Page) 2-2). [範囲指定 (## - ##)をハイフンで分割した 1 番目の要素] 1-3). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:Documentation/def:DocumentRef/def:PDFPageRef (Last Page) 2-3). [範囲指定 (## - ##)をハイフンで分割した 2 番目の要素]
	REFDOCID (DOCUMENTATION) 2	REFDOCID (DOCUMENTATION)と同様
	REFDOCID (DOCUMENTATION) 3	REFDOCID (DOCUMENTATION)と同様
	PROGRAMMING STATEMENTS	プログラムコード 1). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:ProgrammingCode/arm:Code 2). [PROGRAMMING STATEMENTS]
	REFDOCID (PGM. STATEMENTS)	プログラムコードの参照文書 1 記載方法： 文書単位の ID / 参照方法 (PhysicalRef / NamedDestination) / 参照先 a) a) 参照方法が PhysicalRef の場合は，ページ番号 or ページ範囲 (範囲の場合は## – ## ハイフンで区切る)．NamedDestination の場合は，ファイル参照名 文書単位の ID： 1). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:ProgrammingCode/def:DocumentRef (leafID) 2). LF.[文書単位の ID] 参照方法： 1). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:ProgrammingCode/def:DocumentRef/def:PDFPageRef (Type) 2). [参照方法] 参照先： 1-1). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:ProgrammingCode/def:DocumentRef/def:PDFPageRef (PageRefs) 2-1). [参照先] 1-2). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:ProgrammingCode/def:DocumentRef/def:PDFPageRef (First Page) 2-2). [範囲指定 (## - ##)をハイフンで分割した 1 番目の要素] 1-3). arm:AnalysisResultDisplays/arm:ResultDisplay/arm:AnalysisResult/arm:ProgrammingCode/def:DocumentRef/def:PDFPageRef (Last Page) 2-3). [範囲指定 (## - ##)をハイフンで分割した 2 番目の要素]
	REFDOCID (PGM. STATEMENTS) 2	REFDOCID (PGM. STATEMENTS)と同様
	REFDOCID (PGM. STATEMENTS) 3	REFDOCID (PGM. STATEMENTS)と同様

プログラム仕様書で定義する define-xml 構成要素は出来る限り省略しており，定数値として定義（表 2）ができる要素や，プログラム仕様書で定義されている情報から自動的に生成できる要素（表 1 の Description 列），プログラム仕様書の作成時に運用レベルで解決できるような要素は，プログラム仕様書には入力箇所を設けていない．

表 2. 定数値として定義可能な項目

Element	Attribute	Constant Value
ODM	xmlns	http://www.cdisc.org/ns/odm/v1.3
	xmlns:def	http://www.cdisc.org/ns/def/v2.0
	xmlns:xlink	http://www.w3.org/1999/xlink
	ODMVersion	1.3.2

	Filetype	snapshot
MetaDataVersion	def:DefineVersio	2.0.0
	def:StandardName	ADaM-IG
	def:StandardVersion	1.0
TranslatedText	xml:lang	en
ItemGroupDef	Purpose	Analysis
RangeCheck	SoftHard	Soft
arm:ProgrammingCode	Context	SAS Version 9.2
MethodDef	Type	Computation (Value Level Metadata 定義の場合のみ)

今後の運用の中で，各試験でほとんど設定情報が無いような項目がある場合は，プログラム仕様書から削除していく予定である．

3. define.xml の作成方法

3.1 SAS ODS markup

SAS ODS markup は，XML や HTML のようにタグに囲われた文書の入出力に関連する機能を有しており，図 1 の SAS dataset から define.xml を作成するような場合に，あらかじめ準備していたタグセット（テンプレート）と，define.xml 構成要素となるデータセットから，define.xml を出力することができる．

```
ODS markup tagset=sample.definexml_v2 file="C¥:xmlout¥define.xml";
/* ODM */
proc print data=WORK.odm;
run;
/* ODM/Study/ItemGroup */
proc print data=WORK.itemgroup;
run;
/* ODM/Study/ItemGroup/ItemRef */
proc print data=WORK.adsl;
run;
ODS markup close;
```

図 2. SAS ODS markup 実行方法

3.2 テンプレートの作成

図 2 で示した実行方法からも分かるように，define.xml を作成するための処理は，全てテンプレートの定義に集約されている．以下に define.xml を作成するためのテンプレートプログラムの例を示す(図 3)．このサンプルは，例示用に ODM，ItemGroupDef，ItemRef Element のみ作成する仕様になっている（図 5)．使用するデータソースは図 4 に示す．

```
proc template;

  define tagset sample.definexml_v2;

    indent = 2;
    map      = '<>&"';
    mapsub = '/&lt;/&gt;/&amp;/&quot;/&apos;/';

    /* store dataset name to DATANAME */
    define event leaf;
      unset $dataname;
      unset $col_names;
      set $dataname prxchange('s/^Data Set ¥w+¥.(¥w+)$/$1/', 1, value);
    end;

    /* store variable names containing input dataset */
    define event colspec_entry;
      set $col_names[] NAME;
    end;
```

```
/*** Row level processing */
define event row; /* <-- define transaction per one record */
  start   :
    eval   $idx 1;
    unset $col_values;
  finish :
    trigger ODMattr     / if cmp(upcase($dataname), 'ODM');
    trigger IGattr      / if cmp(upcase($dataname), 'ITEMGROUP');
    trigger IRattr      / if cmp(upcase($dataname), 'ADSL');
    break;
end;

/*** Data level processing */
define event data; /* <-- define transaction per 1 data (one variable) */
  start   :
    unset $vname;
    set $vname $col_names[$idx];
    set $col_values[$vname] VALUE;
  finish :
    eval $idx $idx+1;
    break;
end;

/*************************************************************************
***** XML version and style sheet
*************************************************************************/
define event XMLversion;
  start   :
    put   '<?xml version="1.0" encoding="UTF-8" ?>' NL;
    put   '<?xml-stylesheet type="text/xsl" href="../stylesheets/define2-0-0.xsl"?>' NL;
  finish :
    break;
end;


/*************************************************************************
***** ODM
*************************************************************************/
define event ODMmeta;
  start   :
    put '<ODM';
  finish :
    put '>' NL;
end;

define event ODMattr;
  unset $attr;
  set $attr $col_values["ITEM"];

  unset $val;
  set $val   $col_values["DESCRIPTION"] ;

  do / if $attr;
    put ' ';
    put $attr '=';
    do / if cmp($attr, 'CreationDateTime');
      put "" DATE 'T' TIME "" ;
    else ;
      putq $val ;
    done;

  else ;
    break;
  done;
end;

/*************************************************************************
***** ODM/Study/MetadataVersion/ItemGroup
*************************************************************************/
define event IGmeta;
  start   :
    ndent;ndent;ndent;
    put '<ItemGroupDef';
  finish :
    put '</ItemGroupDef>' NL;
    ndent;ndent;ndent;
    break;
end;

define event IGattr;
    unset $name;
    unset $val;

    set $name $col_values["NAME"];
    do / if $name;
      put ' IsReferenceData="No" Purpose="Analysis"';
```

```
        put ' OID="IG.'$name'"';
        put ' NAME=' quote($name);
        put ' SASDatasetName=' quote($name);
        put ' def:CommentOID="COM.'$name'"';
        put ' def:ArchiveLocationID="LF.'$name'"';
      else ;
        break;
      done;

      set $val $col_values["REPEATING"];
      do / if $val;
        put ' Repeating=' quote($val);
      done;

      set $val $col_values["CLASS"];
      do / if $val;
        put ' def:Class=' quote($val);
      done;

      set $val $col_values["STRUCTURE"];
      do / if $val;
        put ' def:Structure=' quote($val);
      done;

      put '>' NL; /* Close ItemGroupDef */

      unset $desc;
      set $desc $col_values["DESCRIPTION"];
      do / if $desc;
        trigger Desc;
      done;

      unset $path;
      set $path $col_values["PATH"];
      do / if $path;
        trigger LEAF_TITLE;
      done;
  end;

  /*************************************************************************
  ***** ODM/Study/MetadataVersion/ItemGroup/ItemRef
  *************************************************************************/
  define event IRmeta;
    start :
      ndent;
    finish :
      xdent;
      trigger IGmeta;
      break;
  end;


  define event IRattr;
    unset $var;
    unset $val;

    set $var $col_values["VARIABLE"];
    do / if $var;
      put '<ItemRef ItemOID="IT.' $dataname '.' $var '"';
    else ;
      break;
    done;

    set $val $col_values["ORDER"];
    do / if $val;
      put ' OrderNumber=' quote($val);
    done;

    set $val $col_values["REQ"];
    do / if $val;
      put ' Mandatory=' quote($val);
    done;

    set $val $col_values["KEY"];
    do / if strip($val);
      put ' KeySequence=' quote($val);
    done;

    set $val $col_values["METHTYPE"];
    do / if cmp(upcase($val), 'COMPUTATION');
      put ' MethodOID="MT.' $dataname '.' $var '"';
    done;

    put '/>' NL; /* Close ItemRef */
  end;

  /*************************************************************************
```

```
      ***** Description/TranslatedText
      ************************************************************************/
      define event Desc;
         ndent;
         put '<Description>' NL;
         ndent;
         put '<TranslatedText xml:lang="en">'$desc'</TranslatedText>' NL;
         xdent;
         put '</Description>' NL;
         xdent;
      end;

      /************************************************************************
      ***** def:leaf/def:title
      ************************************************************************/
      define event LEAF_TITLE;
         ndent;
         put '<def:leaf ID="LF.'$name'" xlink:href=' quote($path) '>' NL;
         ndent;
         put '<put:title>'$name'.xpt</def:title>' NL;
         xdent;
         put '</def:leaf>' NL;
         xdent;
      end;

      /* Document Level */
      define event doc;
         start   :
            trigger XMLversion;
         finish :
            break;
      end;

      /* Document Body Level */
      define event table_body;
         start:
            trigger ODMmeta     / if cmp(upcase($dataname), 'ODM'); * <--- input dataset name ;
            trigger IGmeta      / if cmp(upcase($dataname), 'ITEMGROUP'); * <--- input dataset name ;
            trigger IRmeta      / if cmp(upcase($dataname), 'ADSL'); * <--- input dataset name ;

            break;
         finish:
            trigger ODMmeta     / if cmp(upcase($dataname), 'ODM'); * <--- input dataset name ;
            trigger IRmeta      / if cmp(upcase($dataname), 'ADSL'); * <--- input dataset name ;
            break;
      end;

   end; /* end of sample.definexml_v2 */

run;
```

図 3. テンプレートサンプル

ODM

	ITEM	DESCRIPTION
1	xmlns	http://www.cdisc.org/ns/odm/v1.3
2	xmlns:def	http://www.cdisc.org/ns/def/v2.0
3	xmlns:xlink	http://www.w3.org/1999/xlink
4	ODMVersion	1.3.2
5	FileOID	Sample-SUGJ-2014
6	FileType	Snapshot
7	CreationDateTime	
8	Originator	Sample for SUGJ-2014

ITEMGROUP

	NAME	REPEATING	DESCRIPTION	STRUCTURE	CLASS	PATH
1	ADSL	No	Subject-Level Analysis	One record per subject	ADSL	/ADSL.xpt

ADSL

	VARIABLE	REQ	TYPE	LENGTH	DISPFMT	CNTRLTERM	DERIVTYPE	METHTYPE	ORDER	KEY
1	STUDYID	No	text	18			Predecessor		1	1
2	USUBJID	No	text	11			Predecessor		2	2
3	SUBJID	No	text	4			Predecessor		3	
4	SITEID	No	text	3			Predecessor		4	
5	SITEGR1	No	text	3			Derived	Computation	5	
6	ARM	No	text	20		ARM	Predecessor		6	
7	TRT01P	No	text	20		ARM	Predecessor		7	
8	TRT01PN	No	integer	8		ARMN	Assigned		8	
9	TRT01A	No	text	20		ARM	Assigned		9	
10	TRT01AN	No	integer	8		ARMN	Assigned		10	
11	TRTSDT	No	integer	8	date9		Derived	Computation	11	
12	TRTEDT	No	integer	8	date9		Derived	Computation	12	
13	TRTDUR	No	integer	8		ISO601	Derived	Computation	13	
14	AVGDD	No	integer	8			Derived	Computation	14	
15	CUMDOSE	No	integer	8			Derived	Computation	15	
16	AGE	No	integer	8			Predecessor		16	
17	AGEGR1	No	text	5		AGEGR1	Derived	Computation	17	

図 4. データソース

このデータソースは，表 1 で示したプログラム仕様書の構成に従い定義したプログラム仕様の一部を抽出し，define-xml の属性として不足している情報を導出した後ものになっている．

```
<?xml version="1.0" encoding="UTF-8" ?>
<?xml-stylesheet type="text/xsl" href="../stylesheets/define2-0-0.xsl"?>
<ODM xmlns="http://www.cdisc.org/ns/odm/v1.3" xmlns:def="http://www.cdisc.org/ns/def/v2.0"
xmlns:xlink="http://www.w3.org/1999/xlink" ODMVersion="1.3.2" FileOID="Sample-SUGJ-2014" FileType="Snapshot"
CreationDateTime="2014-07-01T22:10:39" Originator="Sample for SUGJ-2014">
    <ItemGroupDef IsReferenceData="No" Purpose="Analysis" OID="IG.ADSL" NAME="ADSL" SASDatasetName="ADSL"
def:CommentOID="COM.ADSL" def:ArchiveLocationID="LF.ADSL" Repeating="No" def:Class="ADSL" def:Structure="One record
per subject">
      <Description>
        <TranslatedText xml:lang="en">Subject-Level Analysis</TranslatedText>
      </Description>
      <def:leaf ID="LF.ADSL" xlink:href="./ADSL.xpt">
        <put:title>ADSL.xpt</def:title>
      </def:leaf>
      <ItemRef ItemOID="IT.ADSL.STUDYID" OrderNumber=" 1" Mandatory="No" KeySequence="1"/>
      <ItemRef ItemOID="IT.ADSL.USUBJID" OrderNumber=" 2" Mandatory="No" KeySequence="2"/>
      <ItemRef ItemOID="IT.ADSL.SUBJID" OrderNumber=" 3" Mandatory="No"/>
      <ItemRef ItemOID="IT.ADSL.SITEID" OrderNumber=" 4" Mandatory="No"/>
      <ItemRef ItemOID="IT.ADSL.SITEGR1" OrderNumber=" 5" Mandatory="No" MethodOID="MT.ADSL.SITEGR1"/>
      <ItemRef ItemOID="IT.ADSL.ARM" OrderNumber=" 6" Mandatory="No"/>
      <ItemRef ItemOID="IT.ADSL.TRT01P" OrderNumber=" 7" Mandatory="No"/>
      <ItemRef ItemOID="IT.ADSL.TRT01PN" OrderNumber=" 8" Mandatory="No"/>
      <ItemRef ItemOID="IT.ADSL.TRT01A" OrderNumber=" 9" Mandatory="No"/>
      <ItemRef ItemOID="IT.ADSL.TRT01AN" OrderNumber="10" Mandatory="No"/>
      <ItemRef ItemOID="IT.ADSL.TRTSDT" OrderNumber="11" Mandatory="No" MethodOID="MT.ADSL.TRTSDT"/>
      <ItemRef ItemOID="IT.ADSL.TRTEDT" OrderNumber="12" Mandatory="No" MethodOID="MT.ADSL.TRTEDT"/>
      <ItemRef ItemOID="IT.ADSL.TRTDUR" OrderNumber="13" Mandatory="No" MethodOID="MT.ADSL.TRTDUR">/
      <ItemRef ItemOID="IT.ADSL.AVGDD" OrderNumber="14" Mandatory="No" MethodOID="MT.ADSL.AVGDD"/>
      <ItemRef ItemOID="IT.ADSL.CUMDOSE" OrderNumber="15" Mandatory="No" MethodOID="MT.ADSL.CUMDOSE"/>
      <ItemRef ItemOID="IT.ADSL.AGE" OrderNumber="16" Mandatory="No"/>
      <ItemRef ItemOID="IT.ADSL.AGEGR1" OrderNumber="I7" Mandatory="No" MethodOID="MT.ADSL.AGEGR1"/>
    </ItemGroupDef>
```

図 5. 実行結果

テンプレートを使うことで，テーブル・レコード・データ単位での開始と終了のタイミングで処理を定義することができるため，XML のタグを閉じたり，define-xml の element を挿入したりといった処理を定義しやすい．

4. まとめ

昨今，define.xml を作成するツールは数多く提供されており，以前に比べ definx.xml を作成するためのハードルは低くなってきた．しかし，プログラム仕様書に定義された情報を define.xml のデータソースとするような場合，運用上の問題や，プログラム仕様書の使いやすさ，define-xml のバージョンアップといったように，define.xml を作成するためのソースが変わるような要因は数多くあり，それらを公開されているツールでカバーするには限界がある．そのため，どういう方法であれ，社内で define.xml を作成するためのオプションは持っておく必要があると考えている．今後，より効率的なテンプレートの定義方法や，define-xml の複数バージョンへの対応できるような define.xml の作成方法を検討していきたい．

参考文献

CDISC http://www.cdisc.org/

PROC MIANALYZEを用いた多重代入法による結果の統合

石田和也・斎藤和宏
株式会社タクミインフォメーションテクノロジー

Combination of Results for Multiple Imputation Using PROC MIANALYZE

Kazuya Ishida, Kazuhiro Saito
Takumi Information Technology Inc.

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

欠損値を含むデータの解析方法の1つに多重補完法(Multiple Imputation)がある。SASではPROC MIで欠損値の補完をした後、結果を統合するためにPROC MIANALYZEを用いる。本発表では、PROC MIANALYZEについて中心にご紹介する。
なお、SASのバージョンは9.4(SAS/STAT 12.3)を使用した。

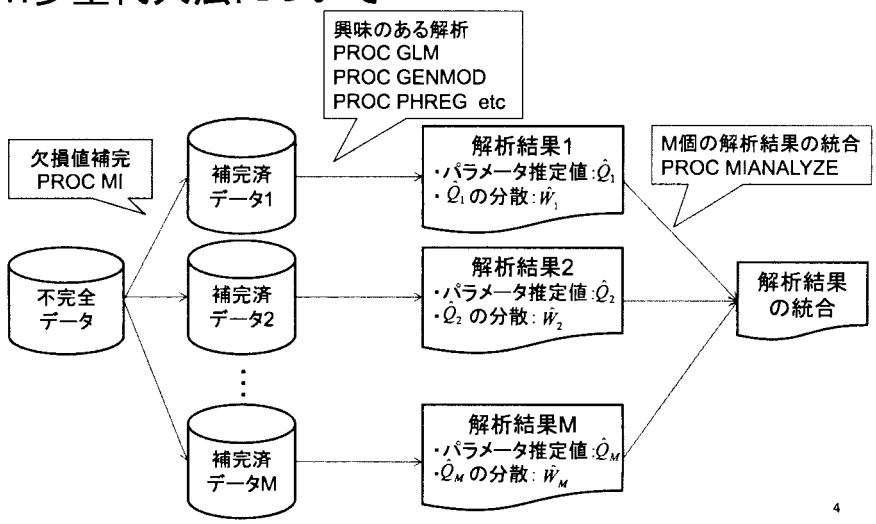
キーワード

PROC MI, PROC MIANALYZE, Multiple Imputation, TYPE3 Test, PROC SQL,

本日の発表構成

1. 多重代入法について
2. 利用事例
3. PROC MIによる欠損値補完
4. 補完後データセットの解析
5. PROC MIANALYZEによる結果の統合
6. 統合後の結果に対する主効果のTYPE3検定について

1.多重代入法について



1.多重代入法について

◆ 補完後データによる解析の統合(Rubin 1987)

➢多重代入法によるパラメータQの推定値 $\overline{Q}$ $\quad \overline{Q} = \frac{1}{M}\sum_{i=1}^{M}\hat{Q}_i$

➢$\overline{Q}$ の分散T $\quad T = \overline{W} + \left(1 + \frac{1}{M}\right)B$

➢ここで $B = \frac{1}{M-1}\sum_{i=1}^{M}(\hat{Q}_i - \overline{Q})^2$ 補完データセット間の分散(推定間のばらつき)

➢ $\overline{W} = \frac{1}{M}\sum_{i=1}^{M}\hat{W}_i$ 補完データセット内の分散

➢多重代入法によるパラメータ推定値の検定 $(H_0 : Q = q_0)$

$$\frac{\overline{Q} - q_0}{\sqrt{T}} \sim t(\nu) \qquad \nu = (M-1)\left[1 + \frac{\overline{W}}{\left(1+\frac{1}{M}\right)B}\right]^2$$

Mが大きい(=補完データセット数が多い)
→自由度νが大きくなる
→検定統計量は漸近的に正規分布にしたがう。

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

2.利用事例

◆ 低体重出生児に対するリスクの解析

➢1986年マサチューセッツ州、スプリングフィールドにあるBaystate Medical Centerにおいて集められたデータの一部(**データに欠損がない、完全データ**)

変数名	ラベル	値	備考
LOW	出生体重が2.5kg を下回るか否か (2.5kgを下回った場合、低体重児とする)	0=低体重児ではない 1=低体重児	目的変数
AGE	母親の年齢(歳)		
LWT	最終月経期間における母親の体重(ポンド)		
RACE	母親の人種	1 = 白人, 2 = 黒人 3 = その他	
SMOKE	妊娠期間の喫煙の有無	0=なし, 1=あり	
PTD	過去の早産の有無	0=なし, 1=あり	
HT	高血圧症の有無	0=なし, 1=あり	
UI	子宮炎症の有無	0=なし, 1=あり	
FTV	妊娠後最初の3か月間に医師の診断を受けた回数	0=0回, 1=1回, 2+=2回以上	

出典:Hosmer, D.W. and Lemeshow, S.(1989).Applied Logistic Regression. Wiley Series in Probability and Statistics.

2.利用事例

◆ 完全データに対するPROC LOGISTICによる解析

```
proc logistic data = LBWI ;
  class race(ref="white") smoke(ref="No")  ptd(ref="No")
        ht(ref="No")       ui(ref="No")       ftv(ref="0") / param=ref  ;
   model low(event="1") = age lwt race smoke ptd ht ui  ;
run ;
```

◆ 完全データに対するPROC GENMODによる解析

```
proc genmod data = LBWI  descending ;
  class race(ref="white") smoke(ref="No")  ptd(ref="No")
        ht(ref="No")       ui(ref="No")       ftv(ref="0") / param=ref  ;
   model low = age lwt race smoke ptd ht ui  / link = logit d = binomial  ;
run ;
```

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

2.利用事例

◆ 完全データに対するPROC LOGISTICによる解析結果
（PROC GENMODの結果は同じなので割愛）

最尤推定値の分析

パラメータ		自由度	推定値	標準誤差	Wald カイ 2 乗	Pr > ChiSq
Intercept		1	0.6369	1.2303	0.2680	0.6047
age		1	-0.0377	0.0378	0.9968	0.3181
lwt		1	-0.0149	0.00704	4.4851	0.0342
race	black	1	1.2127	0.5325	5.1870	0.0228
race	other	1	0.8041	0.4484	3.2153	0.0730
smoke	Yes	1	0.8464	0.4081	4.3020	0.0381
ptd	Yes	1	1.2218	0.4630	6.9626	0.0083
ht	Yes	1	1.8387	0.7033	6.8359	0.0089
ui	Yes	1	0.7111	0.4631	2.3578	0.1247

2.利用事例

◆ 欠損を発生させた際のPROC LOGISTICによる解析結果

最尤推定値の分析

パラメータ		自由度	推定値	標準誤差	Wald カイ 2 乗	Pr > ChiSq
Intercept		1	0.5787	1.2983	0.1987	0.6558
age		1	-0.0349	0.0399	0.7656	0.3816
lwt		1	-0.0145	0.00752	3.7261	0.0536
race	black	1	1.2646	0.5622	5.0600	0.0245
race	other	1	0.7702	0.4579	2.8297	0.0925
smoke	Yes	1	0.7892	0.4209	3.5149	0.0608
ptd	Yes	1	1.0294	0.4992	4.2522	0.0392
ht	Yes	1	2.0163	0.7435	7.3553	0.0067
ui	Yes	1	0.8350	0.4857	2.9564	0.0855

➢欠損を含むオブザベーションが解析から除外されているため・・・
  ➢完全データでは有意であったLWT(最終月経期間における母親の体重)、SMOKE(喫煙有無)が有意ではない。
  ➢全体的にパラメータの標準誤差が大きくなっている
➢解析の精度が下がっていることが分かる

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

3.PROC MIによる欠損値補完

◆ PROC MIを用いた欠損値補完のプログラム例

```
proc mi data = LBWI_miss out = LBWI_MI noprint
                    seed = 123456 nimpute = 20 ;
 class race ftv ht  ;
 fcs discrim(race / classeffects=include)  logistic(ftv)  logistic(ht)  reg(lwt);
 var race ftv ht lwt age ;
run ;
```

補完データセット数Mの設定

判別関数による多重補完(名義尺度)

ロジスティック回帰による多重補完(順序尺度)

線形回帰による多重補完(連続変数)

➢SAS9.3より追加されたFCSステートメントにより、Fully Conditional Specificationによる補完が可能になった。(SAS9.3は評価版)
→RACEのような名義尺度についてもPROC MIによる多重補完が可能となった。

4.補完後データセットによる解析

◆ PROC LOGISTICによる解析

```
proc logistic data = LBWI_MI outest=Log_Param covout  ;
  by _imputation_  ;
  class race(ref="white") smoke(ref="No")  ptd(ref="No")
        ht(ref="No")       ui(ref="No")       ftv(ref="0") / param=ref  ;
   model low(event="1") = age lwt race smoke ptd ht ui  ;
run ;
```

➢補完データセットごとの解析
➢パラメータとその分散共分散行列のデータセット化

◆ PROC GENMODによる解析

```
ods output ParameterEstimates=Gen_Param  COVB=Gen_Cov
           Parminfo=Gen_Info ;
proc genmod data = LBWI_MI  descending ;
  by _imputation_  ;
  class race(ref="white") smoke(ref="No")  ptd(ref="No")
        ht(ref="No")       ui(ref="No")       ftv(ref="0") / param=ref  ;
   model low = age lwt race smoke ptd ht ui  / link = logit d = binomial  ;
run ;
```


5.PROC MIANALYZEによる結果の統合

◆ PROC LOGISTICによる結果の統合

```
proc mianalyze data =  Log_Param  ;
   modeleffect  Intercept age lwt raceblack raceother smokeYes ptdYes
                htYes uiYes ;
run ;
```

➢MODELEFFECTステートメントにおいて、カテゴリ変数を、「変数名+水準」の形で指定する。

➢OUTEST=オプションを使用することができるREGプロシジャ、PHREGプロシジャなどは、同じ指定方法となる。

パラメータと分散共分散行列が1つのデータセットにまとめられているので、PROC MIANALYZEの指定方法は他プロシジャと比べて容易

VIEWTABLE: Work.Log_param (Parameter Estimates and Covariance Matrix)

	Imputation Number	Link function	Type of Statistics	収束状態	Row Names for Parameter Estimates and Covariance Matrix	Intercept: low=0	母親
1	1	LOGIT	PARMS	0 Converged	low	0.736747535	****
2	1	LOGIT	COV	0 Converged	Intercept	1.5030304673	****
3	1	LOGIT	COV	0 Converged	age	-0.027282298	****
4	1	LOGIT	COV	0 Converged	lwt	-0.005628792	****
5	1	LOGIT	COV	0 Converged	raceblack	-0.041931557	****
6	1	LOGIT	COV	0 Converged	raceother	-0.215891358	****
7	1	LOGIT	COV	0 Converged	smokeYes	-0.124731095	****
8	1	LOGIT	COV	0 Converged	ptdYes	0.8205272132	****
9	1	LOGIT	COV	0 Converged	htYes	0.095609564	****
10	1	LOGIT	COV	0 Converged	uiYes	-0.081275363	****
11	2	LOGIT	PARMS	0 Converged	low	0.5615715911	****
12	2	LOGIT	COV	0 Converged	Intercept	1.5589405521	****

5.PROC MIANALYZEによる結果の統合

◆ PROC MIANALYZEによる出力結果

Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
intercept	0.701462	1.246553	-1.74198	3.14490	12163
age	-0.041679	0.037776	-0.11572	0.03236	144417
lwt	-0.014656	0.007191	-0.02875	-0.00056	9099.1
raceblack	1.247608	0.537399	0.19428	2.30094	27594
raceother	0.774883	0.451135	-0.10953	1.65930	5271.9
smokeYes	0.835670	0.405126	0.04159	1.62975	20539
ptdYes	1.210760	0.465283	0.29881	2.12271	89009
htYes	1.758649	0.743711	0.29954	3.21776	1203.7
uiYes	0.694328	0.466967	-0.22093	1.60958	75355

Parameter Estimates

Parameter	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
intercept	0.375152	1.181208	0	0.56	0.5736
age	-0.050861	-0.035429	0	-1.10	0.2699
lwt	-0.018433	-0.012926	0	-2.04	0.0416
raceblack	1.085831	1.374088	0	2.32	0.0203
raceother	0.489834	0.968404	0	1.72	0.0859
smokeYes	0.728863	0.961614	0	2.06	0.0391
ptdYes	1.074406	1.297235	0	2.60	0.0093
htYes	1.259269	2.086117	0	2.36	0.0182
uiYes	0.584834	0.772848	0	1.49	0.1370

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

5.PROC MIANALYZEによる結果の統合

◆ MODELEFFECTステートメントの自動化

➢パラメータデータセットからマクロ変数を作成する

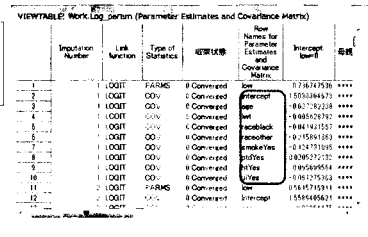
```
proc sql noprint ;
  select _NAME_
  into : modeleffect separated by ' '
  from Log_Param
  where _imputation_=1 and _TYPE_="COV" ;
quit ;
```

SELECT句で指定した変数を
マクロ変数化する。その際・・・・

・変数を横に並べる。
・変数の間はSEPARATED BY で
指定したデリミタで区切る
（左記は半角ブランクで区切る場合）

➢SASログより

```
%put modeleffect = &modeleffect ;
modeleffect = Intercept age lwt raceblack raceother
smokeYes ptdYes htYes uiYes
```



➢マクロ変数を用いると・・・・

```
proc mianalyze data = Log_Param ;
  modeleffect &modeleffect ;
run ;
```

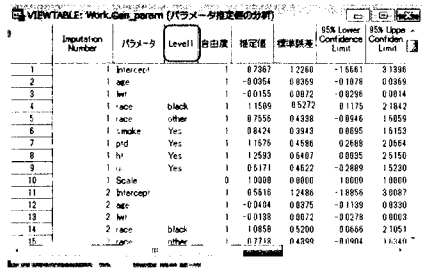
5.PROC MIANALYZEによる結果の統合

◆ PROC GENMODによる結果の統合

```
proc mianalyze parms(classvar=level) = Gen_Param  covb = Gen_COV
                                      parminfo = Gen_info ;
  class race ;
  modeleffect intercept age lwt race  smoke ptd  ht  ui ;
run ;
```

➢CLASSVAR=オプション
- ➢FULL, LEVEL, CLASSVALの中から適切なものを選択。（デフォルトはFULL）
- ➢PROC GENMODの場合、パラメータデータセットのカテゴリ水準の変数名の接頭辞がLEVELなので、LEVELと指定。

➢CLASSステートメントでカテゴリ変数を指定するが、2水準のカテゴリ水準の場合は指定しなくてもよい。



6.統合後の結果に対する主効果のTYPE3検定について

◆ Multivariate Inference（Rubin 1987 , Schafer 1997）

➢多重代入法によるパラメータベクトル Qの推定値 $\overline{\mathbf{Q}}$ $\quad \overline{\mathbf{Q}} = \frac{1}{M}\sum_{i=1}^{M}\hat{\mathbf{Q}}_i$

Ex) Q：人種を表すパラメータベクトル

$\mathbf{Q} = \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \begin{bmatrix} \text{黒人と白人の平均の差} \\ \text{その他と白人の平均の差} \end{bmatrix}$

➢ $\overline{\mathbf{Q}}$ の分散 $\mathbf{T}_0$ $\quad \mathbf{T}_0 = \overline{\mathbf{W}} + \left(1 + \frac{1}{M}\right)\mathbf{B}$

➢ここで $\mathbf{B} = \frac{1}{M-1}\sum_{i=1}^{M}(\hat{\mathbf{Q}}_i - \overline{\mathbf{Q}})(\hat{\mathbf{Q}}_i - \overline{\mathbf{Q}})'$ 補完データセット間の分散共分散行列

➢ $\overline{\mathbf{W}} = \frac{1}{M}\sum_{i=1}^{M}\hat{\mathbf{W}}_i$ 補完データセット内の分散共分散行列

➢下記のような統計量F_0を考える

$$F_0 = (\mathbf{Q}_0 - \overline{\mathbf{Q}})'\mathbf{T}_0^{-1}(\mathbf{Q}_0 - \overline{\mathbf{Q}})/p \sim F(p,\nu) \quad (H_0 : \mathbf{Q} = \mathbf{Q}_0)$$

$$p : \text{Number of Level}, \ \nu = (M-1)\left(1 + \frac{1}{r}\right)^2$$

where $r = \left(1 + \frac{1}{M}\right) trace(\mathbf{B}\overline{\mathbf{W}}^{-1})/p$

Average relative increase in variable

6.統合後の結果に対する主効果のTYPE3検定について

◆ Multivariate Inference（Rubin 1987 , Schafer 1997）

➢F統計量 $F_0 = (\mathbf{Q}_0 - \overline{\mathbf{Q}})'\mathbf{T}_0^{-1}(\mathbf{Q}_0 - \overline{\mathbf{Q}})/p \sim F(p,\nu) \quad (H_0 : \mathbf{Q} = \mathbf{Q}_0)$

➢問題点
- ➢補完データセット数が少ない（Mが小さい）ときに、補完データセット間の分散共分散行列Bが不安定となる。特に $M \le p$ のときにはBがフルランクとならない。

➢解決方法
- ➢補完データセット間、補完データセット内の分散共分散行列が**比例関係**にあることを仮定する。
- ➢この仮定の下では $\overline{\mathbf{Q}}$ の分散Tは $\mathbf{T} = \overline{\mathbf{W}} + r\overline{\mathbf{W}} = (1+r)\overline{\mathbf{W}}$

比例関係を仮定しない場合は・・・
$\mathbf{T}_0 = \overline{\mathbf{W}} + \left(1 + \frac{1}{M}\right)\mathbf{B}$

➢このTを用いて、F統計量を構築しなおす。

➢F統計量 $F = (\mathbf{Q}_0 - \overline{\mathbf{Q}})'\mathbf{T}^{-1}(\mathbf{Q}_0 - \overline{\mathbf{Q}})/p \sim F(p,\nu_1) \quad (H_0 : \mathbf{Q} = \mathbf{Q}_0)$

◆ PROC MIANALYZEでは
このロジックを採用している

where $\nu_1 = \begin{cases} \frac{1}{2}(p+1)(m-1)\left(1+\frac{1}{r}\right)^2 & \text{if } t = p(m-1) \le 4 \\ 4 + (t-4)\left[1 + \frac{1}{r}\left(1 - \frac{2}{t}\right)\right]^2 & \text{if } t = p(m-1) > 4 \end{cases}$

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

6.統合後の結果に対する主効果のTYPE3検定について

◆ TYPE3検定とは

➢モデルにカテゴリ変数を用いた解析において、カテゴリ変数に含まれるいずれかの水準の間に有意な差があるかどうかを確認する検定。

➢ $H_0 : \mathbf{Q} = \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$ （L：係数行列（カテゴリ変数に対するダミー変数行列）、β：カテゴリ共変量のパラメータベクトル）

➢Type3検定のp値が0.05未満であれば、有意水準5%で当該カテゴリ変数のいずれかの水準の間に有意な差があると主張することができる。

➢PROC LOGISTIC、PROC GENMODなどの統計解析プロシジャでは、CLASSステートメントで指定された共変量がモデルに含まれている場合はデフォルトでTYPE3検定の結果が出力されることが一般的であるが、PROC MIANALYZEではデフォルトでは出力されない。

➢係数行列Lを理解した上で、自らTYPE3検定を行う（＝プログラムを構築する）必要がある。
➢TYPE3検定はPROC MIANALYZEの**TEST**ステートメントで行うことができる。

6.統合後の結果に対する主効果のTYPE3検定について

◆ 係数行列Lの確認

➢以下、カテゴリ変数をReference Codingで扱うこととする。
➢Reference Codingでは係数行列Lは$(N-1)$次の正方行列(N:カテゴリ変数の水準数)

➢今回の数値例において、人種(白人、黒人、その他の3水準)のTYPE3検定を考える。
➢白人をReferenceとすると、係数行列Lは以下のようになる。

$$\mathbf{L}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

1行目→「黒人」と「白人」の水準の差、2行目→「その他」と「白人」の水準の差をそれぞれ表す。

➢人種のカテゴリ変数に対するパラメータ推定値をβ_1(「黒人」と「白人」の水準の差)、β_2(「その他」と「白人」の水準の差)とする。

➢人種のカテゴリについてのTYPE3検定は下記のようになる。

$$H_0:\mathbf{L\beta}=\mathbf{0}$$

$$\mathbf{L\beta}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}=\mathbf{0}$$

$$\Rightarrow \boxed{\beta_1=0 \quad and \quad \beta_2=0} \qquad vs \qquad H_1:\beta_1\neq 0 \quad or \quad \beta_2\neq 0$$

少なくとも1水準について、Reference水準との間に有意な差がある
(TYPE3検定:分散分散的な検定)

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

6.統合後の結果に対する主効果のTYPE3検定について

◆ PROC MIANALYZEのTESTステートメント(PROC LOGISTICの場合)

```
proc mianalyze data = Log_Param ;
  modeleffect Intercept age lwt raceblack raceother smokeYes ptdYes
              htYes uiYes ;
  test raceblack = 0 , raceother = 0 / mult ;
run ;
```

・帰無仮説をカンマ区切りでコーディング
・MULTオプションで多変量に対する検定を実施

補完データセット間、補完データセット内の分散共分散行列が**比例関係**にあることを仮定している。

◆ PROC MIANALYZEによる出力結果(TESTステートメント)

Multivariate Inference
Assuming Proportionality of Between/Within Covariance Matrices

r Avg Relative Increase in Variance	p Num DF	v_1 Den DF	F F for H0: Parameter=Theta0	Pr > F
0.046696	2	15412	3.07	0.0465

6.統合後の結果に対する主効果のTYPE3検定について

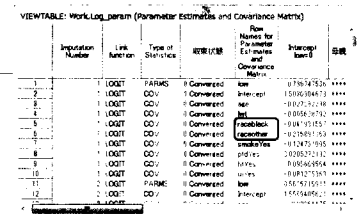
◆ TESTステートメントの自動化

➢パラメータデータセットからマクロ変数を作成する

```
data MakeType3 ;
  set Log_param ;
  where _imputation_ = 1 and  _TYPE_ = "COV"  and  index(_NAME_ , "race") > 0 ;
  Type3 = trim(_NAME_) || "=0" ;   %*TEST Statementのパーツの作成;
run ;

proc sql noprint ;
  select Type3
  into : Type3 separated by ", "
  from MakeType3 ;
quit ;
```

・TEST Statementの構築には、
カンマ区切りのプログラムが必要。
→SEPARATED BY で ", " と指定すればよい



➢マクロ変数を用いると・・・・・

```
proc mianalyze data =  Log_Param  ;
   modeleffect  &modeleffect  ;
   test &MakeType3 / mult ;
run ;
```

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

6.統合後の結果に対する主効果のTYPE3検定について

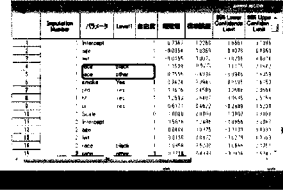
◆ PROC MIANALYZEのTESTステートメント(PROC GENMODの場合)

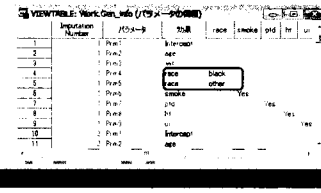
```
proc mianalyze parms(classvar=level) =  Gen_Param  covb = Gen_COV
                                        parminfo = Gen_info ;
  class race ;
  modeleffect intercept age lwt race  smoke ptd  ht  ui ;
  test  race = 0 ;
run ;
```

◆ SASログ

```
WARNING: The TEST statement can not be used when a CLASS statement is specified.
```

➢TESTステートメントと、CLASSステートメントの併用ができない。
→パラメータデータセット(Gen_Param)とパラメータ詳細(Gen_info)データセットの加工が必要





6.統合後の結果に対する主効果のTYPE3検定について

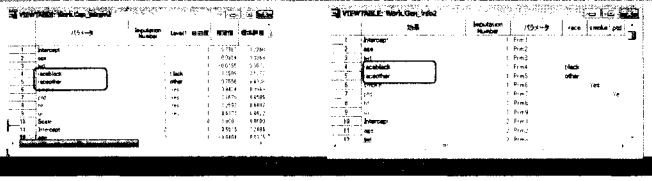
◆ PROC MIANALYZEのTESTステートメント(PROC GENMODの場合)

➢データセットの加工

```
data Gen_Param2 ;
  length Parameter $ 128 ;
  set Gen_Param ;
  if Parameter = "race"  then  Parameter = trim(Parameter) || trim(Level1) ;
run ;

data Gen_info2 ;
  length Effect $ 128 ;
  set Gen_info ;
  if race ^= ""  then  Effect = trim(Effect) || trim(race) ;
run ;
```

カテゴリ変数とその水準を結合して、ユニークな変数値を作成する



医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

6.統合後の結果に対する主効果のTYPE3検定について

◆ 修正後のプログラム(PROC GENMODの場合)

```
proc mianalyze parms =  Gen_Param2  covb = Gen_COV
                        parminfo = Gen_info2 ;
  modeleffect raceblack raceother  ;
  test raceblack = 0 , raceother = 0 / mult ;
run ;
```

◆ PROC MIANALYZEによる出力結果(TESTステートメント)

Multivariate Inference
Assuming Proportionality of Between/Within Covariance Matrices

Avg Relative Increase in Variance	Num DF	Den DF	F for H0: Parameter=Theta0	Pr > F
0.046696	2	15412	3.07	0.0465

➢PROC LOGISTICのときと同じ結果が得られる

まとめ

1. 多重代入法について
   解析の精度が下がることを防ぐため、欠損値を含んだデータに対しては、多重代入法により、欠損値の補完を行った上で解析を行うことが有用である。SASでは、**PROC MIで多重代入法による欠損値補完を、PROC MIANALYZEで結果の統合**を行うことができる。
2. PROC MIANALYZEによる結果の統合
   PROC MIANALYZEに結果の統合を行うためには、補完データセットごとの**解析結果のパラメータデータセットと分散共分散行列のデータセット**が必要となる。これらのデータセットの構造は、統計解析プロシジャにより異なるので、PROC MIANALYZEの指定方法が、使用した統計解析プロシジャにより異なる。
3. 統合後の結果に対する主効果のTYPE3検定について
   モデルにカテゴリ変数を用いた解析では、カテゴリ変数に含まれるいずれかの水準の間に有意な差があるかどうかをTYPE3検定を用いることがある。PROC MIANALYZEでは、**TESTステートメントにより、TYPE3検定**を行うことができる。その際、**補完データセット間、補完データセット内の分散共分散行列が比例関係にあることを仮定している**ので、注意が必要である。

参考文献

◆MI Procedureによる多重代入 SAS ver 9.3における新機能の紹介(2013) / 多田圭佑

◆Multiple Imputation法によるネステッドコントロール研究、ケースコホート研究の解析(2012) / 野間久史・田中司郎・田中佐智子・和泉志津恵

◆ロジスティック回帰分析 SASを利用した統計解析の実際(1996) / 丹後俊郎・山岡和枝・高木春良 著

◆Applied Logistic Regression. Wiley Series in Probability and Statistics (1989) / Hosmer D.W. , Lemeshow S.

◆SAS for Linear Models (1996) / Ramon Littell , Walter W. Stroup, Rudolf Freund

◆Analysis of Incomplete Multivariate Data (1997) / J.L. Schafer

◆SAS/STAT 9.4 User's Guide

ご清聴ありがとうございました

HadoopとSASとの連携テクニック

小林泉
SAS Institute Japan株式会社
ビジネス推進本部アナリティクスプラットフォーム推進

Techniques in SAS on Hadoop

Izumi Kobayashi
Analytics Platform Practice, SAS Institute Japan

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

要旨:

ビッグデータ分析の基盤としてのHadoopの活用が進んでいます。SASはこのような時代の流れを考慮し、SASとHadoopを統合する製品を提供しています。本講演では、ビッグデータをSASで取り扱う機能とテクニックについて取り扱います。主にSASユーザーに向け、Hadoopクラスター上で動作するSASプログラミングテクニックをご紹介し、SASからHadoopの分散処理フレームワークをどのように活用できるかについて解説します。

キーワード: SAS, Hadoop, Map Reduce, Hive, Impala, HDFS, 機械学習, インメモリ, 分散処理

アジェンダ

- Hadoopとは
- HadoopとSAS
- SASデータセットをHDFSに保管する
- Hadoopデータに対してSASから処理を実行する
- SASプログラムをMap Reduceとして分散処理する方法
- Hadoopクラスター上での分散インメモリ処理機能による高速機械学習

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

資料は下記サイトよりダウンロードしてください。

http://www.sascom.jp/campaign/usergroups2014/agenda.php

SASを用いたコピュラに従う擬似乱数の生成

○矢田 真城[1]　　浜田 知久馬[2]

[1] 株式会社 ACRONET データサイエンス本部 生物統計部　[2] 東京理科大学 工学部 経営工学科

Generating pseudo-random numbers from copulas by using SAS system

Shinjo Yada[1]　Chikuma Hamada[2]

[1] Biostatistics Department, ACRONET Corporation

[2] Department of Management Science, Tokyo University of Science

要旨

コピュラ(copula)とは，変数間の依存関係を表す接合関数のことである．複数の変量についてモデル化したい場合，各変量の周辺分布とコピュラとを別々に設定することで，様々な多変量確率分布をあてはめることができる．

医薬品開発において，想定した試験デザインの動作特性を評価するためシミュレーションを行うが，その際，コピュラに従う擬似データを生成しなければならないケースが考えられる．そこで本稿では，代表的なコピュラの乱数生成法について整理し，SAS を用いてこれらの擬似乱数を生成するための具体的な方法を紹介する．

キーワード：擬似乱数，コピュラ，マーシャルオルキン法，COPULA プロシジャ

1. はじめに

医薬品開発において，検討対象となる試験デザインの振舞いをうまく数式で表せない，あるいは数式で表現できてもその解を解析的に求めることが極めて困難な場合，近似的な解を得るためにシミュレーションは威力を発揮する．そのようなシミュレーションでは，大量の乱数を必要とすること，再現性が求められることから，擬似乱数が用いられる[1]．

擬似乱数が従う分布は，試験デザインの評価項目に依存し，コピュラによるモデリングが有効な場合がある．SAS では，擬似乱数を生成させるために RAND 関数が提供されている[2]が，コピュラの場合，RAND 関数で指定できない確率分布が必要とされる．そこで本稿では，正規コピュラ(Normal copula)，スチューデント t コピュラ(Student's t copula)，クレイトンコピュラ(Clayton copula)[3]，フランクコピュラ(Frank copula)[4]，ガンベルコピュラ(Gumbel copula)[5]を取りあげ，各コピュラに従う乱数を生成させる方法について述べる．そして，これら 5 種類のコピュラに対し，SAS を用いて乱数を生成するための方法について具体的に説明する．

2. コピュラについて

m 個の確率変数を $X_1, X_2, \ldots, X_m$ とおき，これらの周辺分布関数を $F_1(x_1), F_2(x_2), \ldots, F_m(x_m)$，同時分布関数を $F(x_1, x_2, \ldots, x_m)$ と表すとき，スクラーの定理(Sklar's theorem)として

$$F(x_1, x_2, \ldots, x_m) = C(F_1(x_1), F_2(x_2), \ldots, F_m(x_m)) \tag{2.1}$$

を満たす関数 C が存在することが知られており，この関数 C がコピュラである[6-8]．これは，関数 C が多次元同時分布とその 1 次元周辺関数を結合する役割を担っていることを示している．よって，m 個の確率変数 $X_1, X_2, \ldots, X_m$ についてモデル化したい場合，個々の周辺分布関数 $F_1(x_1), F_2(x_2), \ldots, F_m(x_m)$ と，これら周辺分布関数の依存関係を表す関数 C，即ちコピュラとを別々に特定することで，$X_1, X_2, \ldots, X_m$ の同時分布関数を構築することができる．

(2.1)において，関数 F が連続関数である場合には C は一意に定まり，任意の $u_i = F_i(x_i)$（ここに i=12,…,m で $u_i \in [0,1]$）に対し

$$C(u_1, u_2, \ldots, u_m) = F(F_1^{-1}(u_1), F_2^{-1}(u_2), \ldots, F_n^{-1}(u_m)) \tag{2.2}$$

と与えられる．このとき C は，各周辺分布が区間[0,1]の一様分布である同時分布関数となる．

変量間の依存関係を行列で表現するコピュラとして，正規コピュラ，スチューデント t コピュラが挙げられる．またそれ以外に，変量間の依存関係を 1 種類のパラメータで表現するコピュラがあり，その代表例がアルキメディアンコピュラ(Archimedean copula)[6-8]である．アルキメディアンコピュラとは，生成素(generator)と呼ばれる関数を用いて表現されるコピュラの総称であり，クレイトンコピュラ，フランクコピュラ，ガンベルコピュラが該当する．

3. 乱数の生成

3.1. アルキメディアンコピュラに従う乱数の生成

区間 (0,1] 上で定義され，0 を含む正の実数値をとる単調減少凸関数 ϕ が

$$\phi(1) = 0 \tag{3.1}$$

を満たすとする．このとき

$$C(u_1, u_2) = \phi^{-1}(\phi(u_1) + \phi(u_2)) \quad \text{ここに } (u_1, u_2) \in (0,1]^2 \tag{3.2}$$

として表現されるコピュラを，2 次元アルキメディアンコピュラと呼び，ϕ を C の生成素という．3 次元以上の場合，(3.1)及び $\lim_{t \to 0} \phi(t) = \infty$ を満たし ϕ^{-1} が完全単調(completely monotone)であるとき

$$C(u_1, u_2, \ldots, u_m) = \phi^{-1}(\phi(u_1) + \phi(u_2) + \ldots + \phi(u_m)), \quad (u_1, u_2, \ldots, u_m) \in (0,1]^m \tag{3.3}$$

として定義される．

アルキメディアンコピュラにおいて，よく用いられる乱数生成法の 1 つがマーシャルオルキン法(Marshall and Olkin method)[9]であり，マーシャルオルキン法を用いた乱数生成法は以下のようになる．手順 1 の $\zeta(\cdot)$ はラプラス変換である．乱数を生成させたいアルキメディアンコピュラに対し，生成素 ϕ の逆関数 $\phi^{-1}(\cdot)$ に対応するような確率分布 $F(\eta)$ を特定し，その確率分布に従う乱数を発生させる必要がある．

手順 1　C の生成素に対し，以下を満たす確率分布 $F(\eta)$ に従う乱数 η を生成させる．

$$\zeta(t)=\int e^{-t\eta}dF(\eta)=\phi^{-1}(t)$$

手順 2　区間[0, 1]の一様分布からの乱数 $x_1,x_2,..,x_m$ を生成させる．

手順 3　$u_i=\zeta(-\eta^{-1}\log(x_i))$　(i=1,2,...,m) を算出する．

3.2. 正規コピュラに従う乱数の生成[10]

平均ベクトル **0**，分散が全て 1 となる分散共分散行列 Σ をもつ多変量正規分布が同時分布の場合，周辺分布は標準正規分布となる．従って，正規コピュラから 1 組の乱数 $u_1, u_2, \ldots, u_m$ を生成するための手順は以下のようになる．

手順 1　所与の相関係数から，分散全て 1 の分散共分散行列を作成する．

手順 2　平均ベクトル **0**，手順 1 で作成した分散共分散行列の多変量正規分布に従う乱数 $z_1, z_2,\ldots, z_m$ を生成させる．

手順 3　$u_i=\Phi(z_i)$　(i=1,2,...,m) を計算する．ここに $\Phi(\cdot)$ は標準正規分布の分布関数である．

3.3. スチューデント tコピュラに従う乱数の生成[10]

自由度 ν，相関行列 $\Sigma(m\times m)$をもつ多変量 t 分布が同時分布の場合，周辺分布は自由度 ν の t 分布となる．従って，スチューデント tコピュラから 1 組の乱数 $u_1,u_2,\ldots,u_m$ を生成するための手順は以下のようになる．

なお，自由度 ν，相関行列 Σ の多変量 t 分布に従う乱数 $x_1,x_2,\cdots,x_m$ は，平均ベクトル **0**，分散共分散行列 Σ の多変量正規分布に従う m 個の乱数 $z_1,z_2,\cdots,z_m$ と，自由度 ν のカイ二乗分布に従う乱数 s とを用いて導出できる．

手順 1　自由度 ν，相関行列 $\Sigma(m\times m)$の多変量 t 分布に従う乱数 $x_1,x_2,\ldots,x_m$ を生成させる．

手順 2　$u_i=t_\nu(x_i)$　(i=1,2,...,m) を計算する．ここに $t_\nu(\cdot)$ は自由度 ν の t 分布の分布関数である．

4. SAS による擬似乱数の生成

コピュラを用いた例として，Yuan and Yin(2009)で用いられた 2 変量生存時間データを参考にした以下のモデルを考える．

T を毒性発現までの時間を表す非負の確率変数とし，T の生存関数としてパラメータ λ_T の指数分布

$$S_T(t_T)=\exp(-\lambda_T t_T) \tag{4.1}$$

を考える．効果発現までの時間 E についても，生存関数としてパラメータ λ_E の指数分布

$$S_E(t_E)=\exp(-\lambda_E t_E) \tag{4.2}$$

を仮定する．更に，毒性発現までの時間 T と効果発現までの時間 E との相関構造として，クレイトンコピュラ(4.3)を想定する．

$$S(t_T, t_E) = \{S_T(t_T)^{-\theta} + S_E(t_E)^{-\theta} - 1\}^{-1/\theta} \tag{4.3}$$

以上に示したモデルに従う生存時間(t_T, t_E)の擬似データは，以下2つのステップにより得ることができる．

<Step1> 想定したコピュラからの擬似乱数を生成する

<Step2> Step1で生成された擬似乱数を変換する

Step1において，想定したコピュラから擬似乱数を生成するにあたり，まずDATAステップを用いた方法を説明する．次に，Ver.9.3よりSAS/ETSに搭載されたCOPULAプロシジャを用いた方法について紹介する．

4.1. DATAステップを用いた擬似乱数の生成

Step1では，クレイトンコピュラ

$$C(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta} \tag{4.4}$$

に従う擬似乱数(u_1,u_2)を生成させる．クレイトンコピュラでは，生成素の逆関数はパラメータθ^{-1}のガンマ分布に従う確率変数のラプラス変換に一致することから，マーシャルオルキン法を用いて，(4.4)で表されるクレイトンコピュラに従う擬似乱数を1組生成させるためのアルゴリズムは，以下のようになる．

手順1　パラメータ$1/\theta$のガンマ分布に従う擬似乱数Zを1つ生成する．

手順2　手順1で生成した擬似乱数Zと独立に，区間[0,1]の一様分布に従う擬似乱数X_1, X_2を生成させる．

手順3　$U_i = (1 - \ln(X_i)/Z)^{-1/\theta}$　(i=1,2) を算出する．

プログラム4.1は，上記の手順1から手順3までを用いて2次元クレイトンコピュラに従う擬似乱数を生成させるためのSASマクロプログラムである．マクロ引数thetaにクレイトンコピュラのパラメータ$\theta(>1)$を，マクロ引数ndrawsに生成する擬似乱数の個数を，マクロ引数outuniformに生成した擬似乱数を保存するためのデータセット名を，それぞれ指定する．

プログラム4.1　DATAステップを用いたクレイトンコピュラの擬似乱数生成

```
%macro mclayton(theta,ndraws,outuniform);
  data &outuniform.;run;
  %do m_i=1 %to &ndraws.;
    data m_wrk ;
     length m_r1 m_r2 m_u 8.; format m_r1 m_r2 m_u best.;
     m_r1=rand('gamma',1/&theta.) ;
     do m_j=1 to 2;
       m_r2=rand('uniform') ;
       m_u =(1-log(m_r2)/m_r1)**((-1)/&theta.);
       output;
     end;
    run;
```

```
    proc transpose data=m_wrk out=m_wrk prefix=m_u;var m_u;run;
    data m_wrk ;set m_wrk;if n(m_u1,m_u2)=2 then SimNo=&m_i.;run;
    data &outuniform.;set &outuniform. m_wrk(keep=SimNo m_u1 m_u2);
      if SimNo^=.;
    run;
    proc datasets library=work nowarn nolist nodetails; delete m_wrk; run; quit;
  %end;
%mend mclayton;
```

Step2では，Step1において得られた擬似乱数(u_1,u_2)を用いて$(S_T^{-1}(u_1), S_E^{-1}(u_2))$と変換することにより，同時分布関数(4.3)に従う擬似乱数を算出する．(4.1), (4.2)より

$$t_T = S_T^{-1}(u_1) = -\ln u_1 / \lambda_T \tag{4.5}$$

$$t_E = S_E^{-1}(u_2) = -\ln u_2 / \lambda_E \tag{4.6}$$

であるから，プログラム4.1により得られた擬似乱数(u_1,u_2)を(4.5), (4.6)へ代入することにより，(4.1)から(4.3)に示したモデルに従う擬似乱数を得ることができる．プログラム4.2は，(4.1), (4.2)においてパラメータをλ_T=1, λ_E=1としたときのSASプログラムである．SASデータセットrandataにt1,t2という変数名でデータが入力される．なお，プログラム4.2では，各症例の観察期間Oは区間$[0,\omega]$の一様分布に従うものとし，i番目のイベント発現までの時間(T_i,E_i)に対し$O_i<T_i$または$O_i<E_i$なら打ち切りとする[12]ように組んでいる．

プログラム4.2　生存時間の擬似乱数を算出するSASプログラム

```
data randata; set unifdata ;
  length lambda_T lambda_E t_1 t_2 8.;
  format lambda_T lambda_E t_1 t_2 best.;
  lambda_T=1; lambda_E=1;
  array unidata(2) m_u1 m_u2;
  array stime(2) t_1 t_2;
  array parm(2) lambda_T lambda_E ;
  do i=1 to 2; stime(i)=round((-1)*(log(1-unidata(i)))/parm(i),1e-12) ;end;
run;
data randata; set randata;
 length w 8.;format w best.; w=7;
 if rand('uniform')*w<t_1 | rand('uniform')*w<t_2 then censor=0; else censor=1;
run;
```

ここでは，2つの生存時間を表す確率変数TとEの依存構造として，クレイトンコピュラ(4.4)を想定したが，フランクコピュラ，ガンベルコピュラは，(3.2)を用いて統一的に表現でき，よってこれらのコピュラに

従う擬似乱数は，クレイトンコピュラと同様の手順で生成することができる．ただし以下に示すとおり，生成アルゴリズムはやや複雑になる．

・フランクコピュラの擬似乱数生成

フランクコピュラ

$$C(u_1,u_2,...,u_m)=\phi^{-1}(\phi(u_1)+\phi(u_2)+\cdots+\phi(u_m))$$

$$=-\frac{1}{\theta}\log\left[1+\frac{\prod_{i=1}^{m}(\exp(-\theta u_i)-1)}{(e^{-\theta}-1)^{m-1}}\right] \tag{4.7}$$

において生成素の逆関数は，パラメータ $\beta=1-e^{-\theta}$ の対数級数分布に従う確率変数のラプラス変換に一致する．よってフランクコピュラの乱数を生成するためには，パラメータ $\beta=1-e^{-\theta}$ の対数級数分布に従う乱数が必要となる．しかし，SAS の RAND 関数では対数級数分布に従う擬似乱数は生成させることができない．逆関数法を用いて擬似乱数を生成させる方法もあるが，ここでは，区間[0,1]の一様分布に従う確率変数 U とパラメータ 1 の指数分布に従う確率変数 V が互いに独立であるとき，

$$X=\left\lfloor -\frac{V}{\ln(1-(1-\beta)^{U})}\right\rfloor+1 \quad (\text{ここに}\lfloor\cdot\rfloor\text{は}\cdot\text{を超えない最大の整数値}) \tag{4.8}$$

によって定義される確率変数 X が対数級数分布に従うことを用いた[13]．この性質を用いれば，区間[0,1]の一様乱数 U とパラメータ 1 の指数分布に従う乱数 V を互いに独立に生成させ，(4.8)に代入することで，対数級数分布に従う乱数を得ることができる．従って，マーシャルオルキン法を用いて，2 次元フランクコピュラに従う擬似乱数を 1 組生成させるためのアルゴリズムは，以下のようになる．

手順 1　以下の手順にて対数級数分布に従う擬似乱数 Z を生成する．

手順 1-1　区間[0,1]の一様乱数 U とパラメータ 1 の指数分布に従う擬似乱数 V とを互いに独立に生成する．

手順 1-2　$Z=\left\lfloor -\frac{V}{\ln(1-(e^{-\theta})^{U})}\right\rfloor+1$ を算出する．

手順 2　手順 1 で生成した擬似乱数 Z とは独立に，区間[0,1]の一様分布に従う擬似乱数 X_1,X_2 を生成させる．

手順 3　$U_i=-\log\{1+\exp(Z^{-1}\ln(X_i))(e^{-\theta}-1)\}/\theta$ （i=1,2）を算出する．

・ガンベルコピュラの擬似乱数生成

ガンベルコピュラ

$$C(u_1,u_2,...,u_m)=\phi^{-1}(\phi(u_1)+\phi(u_2)+\cdots+\phi(u_m))$$

$$=\exp\left\{-\left(\sum_{i=1}^{m}(-\log u_i)^{\theta}\right)^{1/\theta}\right\} \tag{4.9}$$

において，生成素の逆関数は，特性指数(characteristic exponent)が θ^{-1}，歪度パラメータが 1，尺度パラメータ

が$[\cos(\pi/2\theta)]^{\theta}$，位置パラメータが 0 の安定分布 $St(\theta^{-1},1,[\cos(\pi/2\theta)]^{\theta},0)$に従う確率変数のラプラス変換に一致する．従って，マーシャルオルキン法を用いて，2 次元ガンベルコピュラに従う擬似乱数を 1 組生成させるためのアルゴリズムは，以下のようになる．

手順 1　安定分布 $St(\theta^{-1},1,[\cos(\pi/2\theta)]^{\theta},0)$に従う擬似乱数 Z を生成する．
手順 2　手順 1 で生成した擬似乱数 Z とは独立に，区間[0,1]の一様分布に従う擬似乱数 X_1,X_2 を生成させる．
手順 3　$U_i=\exp(-(-\ln(X_i)/Z)^{1/\theta})$　$(i=1,2)$ を算出する．

これを SAS で実装する場合，問題となるのは安定分布に従う擬似乱数の生成である．SAS の RAND 関数では（一部の特殊な場合を除いて）安定分布に従う擬似乱数を生成することができない．そこで，Weron のアルゴリズム[14]により標準安定分布 $St(\alpha,\beta,1,0)$に従う擬似乱数を生成させ，それを用いて任意のパラメータに対応する擬似乱数を作り出すことにした．

Step1　区間$[-\pi/2,\pi/2]$の一様乱数 U を生成する．
Step2　一様乱数 U とは独立にパラメータ 1 の指数分布に従う擬似乱数 V を生成する．
Step3　標準安定分布 $St(\alpha,\beta,1,0)$に従う擬似乱数 X を算出する．

1) $\alpha\neq 1$ のとき

$$X=S_{\alpha,\beta}\frac{\sin(\alpha(U+B_{\alpha,\beta}))}{(\cos(U))^{\frac{1}{\alpha}}}\left(\frac{\cos(U-\alpha(U+B_{\alpha,\beta}))}{V}\right)^{\frac{1-\alpha}{\alpha}}$$

ここに

$$S_{\alpha,\beta}=\left(1+\beta^2\tan^2\left(\frac{\pi\alpha}{2}\right)\right)^{\frac{1}{2\alpha}},\quad B_{\alpha,\beta}=\frac{\arctan\left(\beta\tan\left(\frac{\pi\alpha}{2}\right)\right)}{\alpha}$$

2) $\alpha=1$ のとき

$$X=\frac{2}{\pi}\left[\left(\frac{\pi}{2}+\beta U\right)\tan U-\beta\log\left(\frac{\frac{\pi}{2}V\cos U}{\frac{\pi}{2}+\beta U}\right)\right]$$

上記アルゴリズムにより得られた標準安定分布 $St(\alpha,\beta,1,0)$に従う擬似乱数 X に対し

$$Y=\begin{cases}\sigma X+\mu & (\alpha\neq 1)\\ \sigma X+2\beta\sigma\log\sigma/\pi+\mu & (\alpha=1)\end{cases}\qquad(4.10)$$

とすれば，特性指数 $\alpha\,(0<\alpha\leq 2)$，歪度パラメータ $\beta\,(-1\leq\beta\leq 1)$，尺度パラメータ σ，位置パラメータ μ の安定分布 $St(\alpha,\beta,\sigma,\mu)$に従う擬似乱数となる[15]．

付録 A に，DATA ステップを用いて，クレイトンコピュラ，ガンベルコピュラ，フランクコピュラ，正規コピュラ，スチューデント t コピュラの擬似乱数を生成するプログラムを示した．少し補足すると，正規コ

ピュラ及びスチューデント t コピュラの場合，3.2 節，3.3 節に示したように，多変量正規分布の乱数が必要となる．多変量正規分布に従う乱数の生成は，分散共分散行列を Cholesky 分解することに基づいて行うことができる[5]ため，付録 A では MIXED プロシジャを用いて Cholesky 分解を行い，多変量正規分布に従う擬似乱数を生成させている[16].

4.2. COPULA プロシジャを用いた擬似乱数の生成

SAS では，Ver.9.3 より SAS/ETS に COPULA プロシジャが Experimental として搭載され，正規コピュラ，スチューデント t コピュラ，クレイトンコピュラ，フランクコピュラ，ガンベルコピュラに従う擬似乱数の生成が可能となった[10]．(4.4)で示されるクレイトンコピュラに従う擬似乱数を，COPULA プロシジャを用いて生成するための SAS プログラムを以下に示す．

プログラム 4.3　COPULA プロシジャを用いたクレイトンコピュラの乱数生成

```
proc copula;
   var U1-U2 ;
   define COP clayton (theta=8) ;
   simulate COP/ ndraws=1000 seed=0724 outuniform=unifdata;
run;
```

ユーザーが指定したコピュラに従う擬似乱数を生成することが目的であるので，PROC COPULA ステートメントにおいて，DATA=の指定は不要である．生成する擬似乱数の変数名は，VAR ステートメントに記載する．DEFINE ステートメントにおいて，クレイトンコピュラを表すキーワードである CLAYTON と指定し，あわせてパラメータ値をオプション THETA=で指定する．また，SIMULATE ステートメントにて使用するために，このコピュラの名前を“COP”とした．生成させる擬似乱数に関する指定は SIMULATE ステートメントで行う．SIMULATE ステートメントにおいて，DEFINE ステートメントにおいて定義したコピュラの名前を記述した後，オプション NDRAWS=で生成する擬似乱数の個数を，オプション SEED=で擬似乱数を生成するためのシードを，それぞれ指定する．生成した擬似乱数を保存するための SAS データセット名は，オプション OUTUNIFORM=で記述する．

プログラム 4.3 より，クレイトンコピュラに従う擬似乱数が生成されれば，あとは DATA ステップでの生成と同様に，得られた擬似乱数(u_1,u_2)を用いて$(S_T^{-1}(u_1), S_E^{-1}(u_2))$と変換することにより，同時分布関数(4.3)に従うデータを算出すればよい．なお，2 つの生存時間を表す確率変数 T と E の依存構造として，正規コピュラ，スチューデント t コピュラ，ガンベルコピュラ，フランクコピュラを考えるのであれば，プログラム 4.3 において，DEFINE ステートメントの CLAYTON の代わりに各コピュラを表すキーワード及びパラメータを指定する．ただし，正規コピュラ，スチューデント t コピュラにおいては，分散共分散行列の各成分をもつ SAS データセットを予め用意した上で，DEFINE ステートメントのオプション CORR=にて指定する必要がある．プログラム 4.4 に，ケンドールの τ を 0.8 としたときの正規コピュラに従う擬似乱数を生成するための SAS プログラムを示した．2 変量正規コピュラにおいて，ケンドールの τ と相関係数 ρ とには

$$\tau = (2/\pi)\arcsin\rho \tag{4.11}$$

との関係が成り立つことを用い，DATA ステップで対応する分散共分散行列を用意した上で，COPULA プロシジャにより擬似乱数を生成している．

プログラム 4.4　COPULA プロシジャを用いた正規コピュラの擬似乱数生成

```
data corr;
   length U1 U2 rho pi 8.; format U1 U2 rho pi best.;
   keep U1 U2;
   pi=constant('pi'); rho=sin(0.8*pi/2);
   U1=1;   U2=rho;output;
   U1=rho; U2=1 ; output;
run;
proc copula;
   var U1-U2 ;
   define COP NORMAL(cor=corr) ;
   simulate COP/ ndraws=1000 seed=0724 outuniform=unifdata;
run;
```

5. おわりに

本稿では，SAS を用いてコピュラに従う乱数を生成させる方法について紹介した．コピュラを取り扱うプロシジャとして，SAS Ver.9.3 より SAS/ETS に COPULA プロシジャが搭載され，計 5 種類のコピュラに従う乱数を簡単に生成させることができる．また，当該プロシジャを使用できない環境下で乱数を生成させる方法として，DATA ステップを用いたマクロプログラムを示した．

COPULA プロシジャでは，解析対象となるデータにコピュラをあてはめたときのパラメータを推定することも可能である．現段階では Experimental の扱いであるが，完全にサポートされれば更に利用価値が高まるものと期待される．コピュラを用いたシミュレーション実験あるいはデータ解析を行う際に，本稿がその一助になれば望外の喜びである．

参考文献

[1] 岩崎学(2005). 統計的データ解析のための数値計算法入門. 朝倉書店.

[2] 魚住龍史，浜田知久馬(2013). RAND 関数による擬似乱数の生成. SAS ユーザー総会論文集 2013, 325-333.

[3] Clayton,D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65, 141–151.

[4] Frank,M.J. (1979). On the simultaneous associativity of F(x,y) and x+y-F(x,y). Aequationes Mathematicae 19,

194–226.

[5] Hougaard, P. (1986). A class of multivariate failure time distributions. Biometrika 73, 671–678.

[6] Nelson, R.B. (2006). An Introduction to Copulas. New York: Springer.

[7] 塚原英敦(2012).「 接合分布関数(コピュラ)の理論と応用」. 北川源四郎, 竹村彰通(編). 21 世紀の統計科学 <Vol.III> 数理・計算の統計科学. 東京大学出版会増補 HP 版.
http://park.itc.u-tokyo.ac.jp/atstat/jss75shunen/Vol3.pdf(最終閲覧日: 2014 年 6 月 1 日).

[8] 戸坂凡展, 吉羽要直(2005) .「 コピュラの金融実務での具体的な活用方法の解説」. 日本銀行金融研究所.
http://www.imes.boj.or.jp/japanese/kinyu/2005/kk24-b2-3.pdf(最終閲覧日: 2014 年 6 月 1 日).

[9] Marshall,A.W. and Olkin,I. (1988). Families of Multivariate Distributions. Journal of the American Statistical Association 83, 834–841.

[10] SAS Institute Inc. (2011). SAS/ETS 9.3 User's Guide. The COPULA Procedure. Cary, NC: SAS Institute Inc.

[11] Ying,G. and Yuan,Y. (2009). Bayesian dose finding by jointly modeling toxicity and efficacy as time-to-event outcomes. Journal of the Royal Statistical Society 58, 719-736.

[12] Schemper,M.,Kaider,A.,Wakounig,S. and Heinze,G.. (2013). Estimating the correlation of bivariate failure times under censoring. Statistics in Medicine 32, 4781-4790.

[13] 四辻哲章(2010). 計算機シミュレーションのための確率分布乱数生成法. プレアデス出版.

[14] Weron,R. (1996). On the Chambers-Mallows-Stuck method for simulating skewed stable random variables. Statistics and Probability Letters 28, 165-171.

[15] Clizek,P.,Hardle,W.K. and Weron,R. (2005). Statistical Tools for Finance and Insurance. Springer.

[16] SAS CUSTOMER SUPPORT/TECHNICAL SUPPORT. 多次元正規分布に従う 乱数列を 生成する方法.
http://www.sas.com/offices/asiapacific/japan/service/technical/faq/list/body/stat034.html(最終閲覧日: 2013 年 10 月 19 日).

連絡先

E-mail: s-yada@acronet.jp

付録 A　DATA ステップによるコピュラの乱数生成プログラム

```
/* クレイトンコピュラに従う乱数生成マクロ
  dim.....コピュラの次元
  theta.....コピュラのパラメータ
  seed.....シード
  ndraws.....生成したい乱数の個数
  outuniform.....出力用データセット */
%macro mclaycop(dim,theta,seed,ndraws,outuniform);
   data &outuniform.; keep m_u: SimNo;
    array col{&dim.}; array m_u{&dim.}; call streaminit(&seed.);
    do SimNo=1 to &ndraws.;
       m_r1=rand('gamma',1/&theta.) ;
       do m_j=1 to dim(m_u);
         col{m_j}=rand('uniform') ;
         m_u{m_j}=(1-log(col{m_j})/m_r1)**((-1)/&theta.);
       end;
       output;
     end;
  run;
%mend mclaycop;

/* フランクコピュラに従う乱数生成マクロ
  dim.....コピュラの次元
  theta.....コピュラのパラメータ
  seed.....シード
  ndraws.....生成したい乱数の個数
  outuniform.....出力用データセット */
%macro mfrankcop(dim,theta,seed,ndraws,outuniform);
   data &outuniform.; keep m_u:;
    array col{&dim.}; array m_u{&dim.}; call streaminit(&seed.);
    do SimNo=1 to &ndraws.;
       m_su=rand('uniform') ;
       m_sv=rand('exponential') ;
       m_r1=int((-1)*m_sv/(log(1-(exp((-1)*&theta.))**m_su)))+1;
       do m_j=1 to dim(m_u);
          col{m_j}=rand('uniform');
```

```
        m_u{m_j}=( log( 1+ exp(log(col{m_j})/m_r1)*(exp((-1)*&theta.)-1) ) )*((-1)/&theta.);
      end;
      output;
    end;
  run;
%mend mfrankcop;

/* ガンベルコピュラに従う乱数生成マクロ
 dim.....コピュラの次元
 theta.....コピュラのパラメータ
 seed.....シード
 ndraws.....生成したい乱数の個数
 outuniform.....出力用データセット */
%macro mgumbelcop(dim,theta,seed,ndraws,outuniform);
  data &outuniform.; keep m_u:;
   array col{&dim.}; array m_u{&dim.};
   call streaminit(&seed.);
   do SimNo=1 to &ndraws.;
     length m_xu m_v pi m_x m_s m_b m_r1 8.; format m_xu m_v pi m_x m_s m_b m_r1 best.;
     pi=constant('pi');
     m_gamma=(cos(pi/(2*&theta.)))**&theta.; m_alpha=round(1/&theta.,1e-12); m_beta =1;
     if  m_alpha^=1 then do;
      m_xu=rand('uniform')*pi-pi/2;
      m_v =rand('exponential');
      m_b =atan(m_beta*tan(pi*m_alpha/2));
      m_s =(1+(m_beta*tan(pi*m_alpha/2))**2)**(1/(2*m_alpha));
      m_x =m_s*(sin(m_alpha*m_xu+m_b)/(cos(m_xu)**(1/m_alpha)))
      *( (cos(m_xu-m_alpha*m_xu-m_b)/m_v)**((1-m_alpha)/m_alpha) );
      m_r1=round(m_gamma*m_x,1e-12);
     end;
     else if m_alpha=1 then do;
      m_xu=rand('uniform')*pi-pi/2;
      m_v =rand('exponential');
      m_x =(2/pi)*((pi/2+m_beta*m_xu)*tan(m_xu)-m_beta*log(((pi/2)*m_v*cos(m_xu))/(pi/2+m_beta*m_xu)));
      m_r1=round(m_gamma*m_x+(2/pi)*m_beta*m_gamma*log(m_gamma),1e-12);
     end;
```

```
      do m_j=1 to dim(m_u);
       col{m_j}=rand('uniform') ;
       m_u{m_j}=exp( (-1)*(-log(col{m_j})/m_r1)**(1/&theta.) );
      end;
      output;
     end;
    run;
%mend mgumbelcop;

/* 正規コピュラに従う乱数生成マクロ
 dim.....コピュラの次元
 cor.....相関行列(SAS データセット名). 変数名は colX でないとエラーになる.
 seed.....シード
 ndraws.....生成したい乱数の個数
 outuniform.....出力用データセット */
%macro mnormcop(dim,cor,seed,ndraws,outuniform);
  data m_wrk;
    call streaminit(&seed.);
    _TYPE_="SCORE"; _MODEL_="col"; mean=1; array col{&dim.};
    do num=1 to &ndraws.;do i=1 to dim(col);col{i}=rand("NORMAL");end;output;end;
    drop i;
  run;
  data m_cov ;set &cor.;length mean 8.;format mean best.;mean=0;row=_n_;run;
  ods listing close;
  ods output CHOLG=m_cholesky;
  proc mixed data=m_cov;
   class row mean; parms /noiter;model mean=; random row*mean/type=UN gdata=m_cov GC;
  run;
  ods listing;
  proc score data=m_cholesky score=m_wrk out=m_mnorm(keep=col num); by num;var mean col:; run;
  proc transpose data=m_mnorm OUT=m_mnorm(DROP=_NAME_);by num;run;
  data &outuniform.;set m_mnorm;
   array col{&dim.} ; array m_u{&dim.};
   do i=1 to &dim.;do i=1 to dim(m_u);m_u{i}=CDF('NORMAL',col{i});end;end;
   drop i col:;
  run;
```

```
  proc datasets library=work nowarn nolist nodetails; delete m_wrk m_cholesky m_cov m_mnorm; run; quit;
%mend mnormcop;

/* tコピュラに従う乱数生成マクロ
 dim.....コピュラの次元
 cor.....相関行列(SAS データセット名). 変数名はcolXでないとエラーになる.
 df.....tコピュラの自由度
 seed.....シード
 ndraws.....生成したい乱数の個数
 outuniform.....出力用データセット */
 %macro mtcop(dim,cor,df,seed,ndraws,outuniform);
   data m_wrk;
     call streaminit(&seed.);
     _TYPE_="SCORE"; _MODEL_="col"; mean=1; array col{&dim.};
     do num=1 to &ndraws.;do i=1 to dim(col);col{i}=rand("NORMAL");end;output;end;
     drop i;
   run;
   data m_cov ;set &cor.;length mean 8.;format mean best.;mean=0;row=_n_;run;
   ods listing close; ods output CHOLG=m_cholesky;
   proc mixed data=m_cov;
    class row mean; parms /noiter;model mean=; random row*mean/type=UN gdata=m_cov GC;
   run; ods listing;
   proc score data=m_cholesky score=m_wrk out=m_mnorm(keep=col num); by num;var mean col:; run;
   proc transpose data=m_mnorm OUT=m_mnorm(DROP=_NAME_);by num;run;
   data m_wrk ;do num=1 to &ndraws.;output;end;run;
   data m_wrk;set m_wrk;length m_s 8.;format m_s best.; call streaminit(&seed.+num);m_s=rand('chisquare',&df.);run;
   data &outuniform.;
    merge m_mnorm m_wrk;by num;
    array col{&dim.} ; array m_u{&dim.};
    do i=1 to &dim.;do i=1 to dim(m_u);m_u{i}=CDF('T',col{i}*sqrt(&df./m_s),&df.);end;end;
    drop i col:;
   run;
   proc datasets library=work nowarn nolist nodetails; delete m_wrk m_cholesky m_cov m_mnorm; run; quit;
 %mend mtcop;
```

EXCELで数独解答プロセスをリアルタイムで可視化するSASプログラム

SAS Sudoku Program to Visualize its Solving Process in Real Time Using EXCEL

森岡　裕（ナイフィックス株式会社）
Fuad J. Foty (U.S. Census Bureau)
知平　菜美子（株式会社NSD 金融事業本部）
周防　節雄（兵庫県立大学・名誉教授）

要旨

2011年にSASユーザー総会で発表された「数独パズルを解くSASプログラム」(知平・周防2011、周防・知平2011)で用いられた高度なデータハンドリング技法に森岡は影響を受け、その2年後、森岡はSASユーザー総会で、セルオートマトンというマス目状モデルを用いて行うシミュレーションをSASの環境で行い、その結果を逐次EXCEL画面上に表示することによって、リアルタイムで視覚的に理解できる技法に関する論文(森岡2013)を発表した。その論文発表後に、周防が森岡に、両者のプログラムを一つにして、数独解法プログラムの解答プロセスをEXCEL画面上にリアルタイムで可視化したら、一層興味深いSASプログラムになるのではないかと提案し、今回の共同研究が実現した。以前は解答プロセスを分析するには、プログラム実行中にEXCELファイルに逐次保存しておいた途中結果を解答終了後に別途解析することで可能であったが、今回開発した可視化数独解法プログラム(Visualized, Smart and Slim Sudoku Solving System:略称はVisual S[5])には、プログラム実行中に数独を解いている様子がEXCEL画面上に逐次表示される機能を追加した。空白のマス目に数字を埋めていくための補助情報が逐次更新される一方で、マス目に数字が次々と埋められていく様子を、SASのアウトプット画面や結果ビュー画面ではなくて、カラフルなEXCEL画面上で目の当たりにするのは壮観である。数独の解法アルゴリズムの解説は先行論文に譲り、本論文では、解答プロセスをEXCEL画面上に可視化する技法と、データハンドリングにおけるSASの様々テクニックの紹介に重点をおいてVS[5]を紹介する。

1. 序論

1.1 はじめに

2013年のSASユーザー総会で、森岡はセルオートマトンというマス目状モデルを用いて行うシミュレーションをSASで行い、その結果を逐次EXCEL画面上に表示することによって、リアルタイムで視覚的に理解できる技法に関する論文(森岡2013)を発表した。この論文は、その2年前に周防と知平

が発表した「数独パズルを解く SAS プログラム」(知平・周防 2011、周防・知平 2011)に出会ったのがきっかけで、マス目状のパズルを SAS で解いているのを見て、セル状空間と SAS データセットという着想を得て、執筆したものであった。

昨年の森岡論文(森岡 2013)の発表後、数独プログラムの作成者の一人である周防が、数独解法 SAS プログラムに対して、EXCEL 画面上にリアルタイムで途中経過を可視化する機能の追加を提案し、共同研究が始まった。

周防・知平の数独 SAS プログラム(Sudoku Solving System：略称は SSS、「トリプル S」)は再帰的プログラミングを用い、高度なマクロを数多く組み合わせた複雑な構造であり、数独を解く途中経過は解答処理の終了後、途中経過を保存した EXCEL ファイルを辿って解析するしか確認する術がなかった。そのファイルの解析は SAS のパワーユーザーでなければかなり困難であり、周防はその解法プロセスをもっと簡単に見えるようにしたいと常々考えていた。

一方、周防は 2012 年米国フロリダで開催された SAS Global Forum 2012 で数独論文を招待論文として発表し(Suoh 2012)、SAS プログラム(Windows 版)の日本語版だけでなく英語版もネット上に一般公開した。その時フロアで発表を聞いていた Foty（米国センサス局所属）は、その 3 ヶ月後、マクロの数を大幅に少なくして周防・知平のプログラムのスリム化、高速化を図ったプログラム(Linux 版 SAS)を周防に突然送ってきた。周防・知平の SAS プログラムはマクロが大変複雑で、その上再帰的手法を使っているため、第三者がそれを解読して更に改良することは全く予想していなかったので周防は一瞬途惑ったが、反面、その SAS プログラミングの力量に驚き、大いに感激した。周防と Foty はこのニューバージョンを Smart and Slim Sudoku-Solving System (：略称 S^5「Quadruple S」と発音する)と命名して SAS Global Forum2013 に投稿する予定であったが、双方が多忙なため延期となり現在に至っている。

その間に、森岡のセルオートマトンを使った可視化機能追加の構想が持ち上がった。森岡は、その時点の最新の数独プログラム S^5 を Windows 版に変換した後、数独解答処理中の途中結果の SAS データセットを逐次 EXCE 画面上に表示させる機能を追加した SAS プログラムを完成した。このプログラムを本論文で解説する。今回バージョンアップしたシステムは、視覚化する部分を追加したので Visualized, Smart and Slim Sudoku-Solving System (略称 Visual S^5)と命名したが、Foty の S^5 の数独解法アルゴリズム自体には手を一切加えていない。今回の可視化作業に先立ち、森岡も SAS マクロの更なる最適化が出来ないかと検討してみたが、これ以上のコードの効率化は不可能な位複雑で、完成度が高いことが判明し、解法アルゴリズム自体のこれ以上の改良は行えなかった。

本論文では、解答プロセスを可視化するために追加したプログラムの技術的な部分を説明することに絞った。数独プログラムの開発・改良過程や解法アルゴリズムについては参考文献にあげる論文を年代順に読んでいただきたい。

2. Visual S^5 で使用されている技法

2.1 LIBNAME EXCEL と EXCEL 周辺知識について

Visual S^5 について説明する前に、EXCEL ファイルとの入出力処理に利用している LIBNAME EXCEL (EXCEL ファイルへのライブラリ参照)や、その他、いくつかの細かな技法について解説する。

2.1.1 セル範囲の名前の定義

EXCEL には任意のセル範囲に名前を定義する機能がある。

例えば図１の左図において、中央の９個の数字を SUM 関数で合計してセル「R6:C2」に表示する場合の関数式は「=SUM(R2C2:R4C4)」である。

しかし、「数式タブ」→「名前の定義」(図１の中央図)から参照範囲として「R2C2:R4C4」を選択後、例えば「集計エリア」というテキスト(図１の右図)を入力して「OK」すれば、セル範囲「R2C2:R4C4」に名前が定義される。

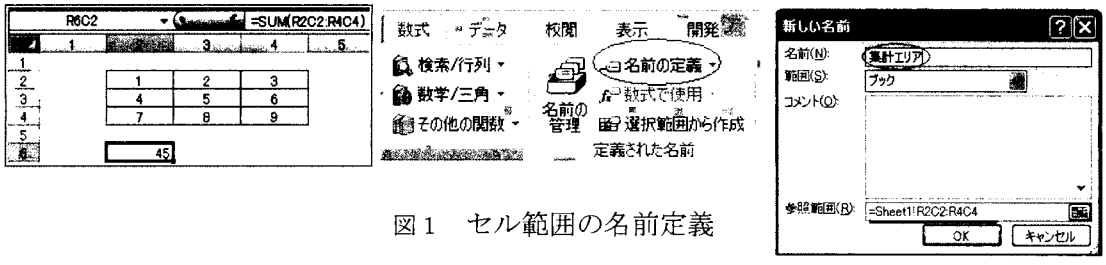


図１　セル範囲の名前定義

その結果、関数式「=SUM(R2C2:R4C4)」の代わりに、「=SUM(集計エリア)」(図２) に置き換えて記述することができるようになる。つまり、セルの「名前の定義」とは、直接セル番地を指定せずに処理を可能とする機能である。

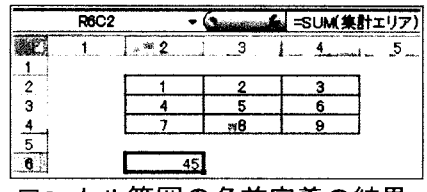


図2 セル範囲の名前定義の結果

2.1.2 LIBNAME EXCEL

LIBNAME 文で EXCEL ファイルのフルパスを指定することによって、EXCEL ファイルがライブラリになり、シートまたは名前の定義が SAS データセットとして認識される。

例えば、図１で例示した EXCEL シートが Book1.xlsx という名前で C ドライブのルートに保存されている場合、LIBNAME 文で以下のように定義すると、普通のシートの場合はシート名の末尾に$マークのついた SAS データセット、名前を定義した範囲はそのままの名前で SAS データセットとして認識される(図３)。

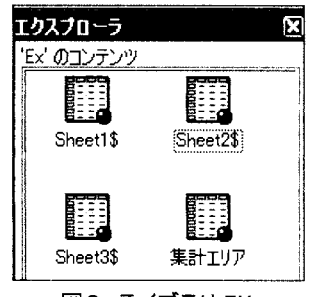


図3 ライブラリ EX

```
libname EX "C:¥Book1.xlsx" header=no scantext=no;
```

図３の SAS データセット「集計エリア」をダブルクリックすると、３行×３列の EXCEL 表の名前の定義が３変数３オブザベーションとなっていることが分かる(図４)。コード中の「header=no」オプションは１行目を変数名とするかどうかのオプションで、no にすると変数名は F1, F2･･･と連番号で自動的に付与される。

VIEWTABLE: Ex.集計エリア

	F1	F2	F3
1	1	2	3
2	4	5	6
3	7	8	9

図4 SAS データセット「集計エリア」

データセットの内容と EXCEL ファイルのセル範囲は同期しているので、どちらか一方の中身が変更されると、それがもう一方に反映される仕組みになっている。

まず初めに、SAS プログラムで EXCEL 表の内容を変更する場合を取り上げる。そのためには MODIFY 文を使用する(**プログラム1**)。ただし、既に存在する EXCEL シートや名前の範囲の内容を SAS

から更新する場合は、LIBNAME 文の実行時に scantext=no オプションを指定しておく必要がある。続いて、①及び②の箇所のように、名前の定義をデータセットとして指定する際には、シングルコーテーションで囲み、n を末尾に付して記述する必要がある。また、名前の定義ではなく、シートを対象とする場合であれば、①の部分は data EX.'Sheet1$'n; のように シート名の末尾に$を付す。②は、①の記述の data を modify に変えるだけでよい。③以降の記述は普通の SAS プログラミングの記述である。

```
data EX.'集計エリア'n; ← ①
  modify EX.'集計エリア'n; ← ②
   F1=F1*2; ← ③
   F2=F2+1;
   F3=0;
run;
```

プログラム1:
Modify 文による
EXCEL 表の加工・出力

実行後に名前の定義「集計エリア」を SAS と EXCEL の環境で見ると、それぞれ図5、図6のようになっている。

VIEWTABLE: Ex.集計エリア

	F1	F2	F3
1	2	3	0
2	8	6	0
3	14	9	0

図5 SAS データセット

2	3	0
8	6	0
14	9	0

図6 EXCEL 表

次に、SAS データセット DS1 (図7)を加工して EXCEL の'集計エリアに出力する場合を考える。プログラム2に示す様に、①の SET 文を data 文と modify 文の間に挿入することで実現できる。

VIEWTABLE: Work.Ds1

	X	Y	Z
1	1	3	5
2	2	4	6
3	1	3	5

図7 SAS データセット DS1

```
data EX.'集計エリア'n;
 set DS1; ← ①
 modify EX.'集計エリア'n;
   F1=X;
   F2=Y;
   F3=Z;
 run;
```

プログラム2:
Modify 文による SAS データセットの加工と EXCEL 出力

LIBNAME EXCEL の使用上の注意点としては、使用する SAS のバージョンと EXCEL のバージョンの組み合わせ、またはその他の実行環境の違いによって実行時の動作が一貫しないことがある。また、対象とする EXCEL ファイルが閉じている状態と開いている状態において、実行時に明らかな動作の差が存在する。現時点で確認できている注意点として重要と思われるものを以下に列挙する。

1) 単一のセルに名前を定義した場合、EXCEL ファイルを閉じている状態ではデータセットとして認識されるが、開いた状態では認識されない
2) 名前で定義されたセル範囲の値をリセットする場合の動作の違い(詳細は次節で説明)

2.1.3 EXCEL VBA の利用

名前の定義で指定されたセル範囲の値をリセット、つまり全て null 値にしたい場合は、対象の EXCEL ファイルが閉じていればプログラム3を実行する。実行後、更に libname EX clear; を実行すればライブラリ指定の解除ができる。EXCEL を閉じて libname 文を使用する場合は、解除せずに開こうとするとエラーが発生するので注意が必要である。EXCEL ファイルを開くと、セルは全て空になっている (図8) が、名前の定義「集計エリア」は依然として存在している。

一方、EXCEL ファイルを開いた状態で同じプログラム3を実行した場合、セル範囲の値がリセットされるところまでは同じであ

```
proc datasets lib=EX nolist;
 delete '集計エリア'n;
run;
quit;
```

プログラム 3:
名前の定義で指定の
セル範囲の値リセット(1)

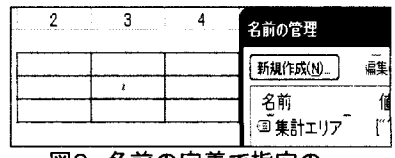


図8 名前の定義で指定の
セル範囲のリセット後

るが、さらに名前の定義が EXCEL ファイルから消失する。実行する度に消えた名前の定義を再度設定するのは面倒なので、EXCEL ファイルを開いた状態で名前の範囲を削除させずに値をリセットしたい場合には、**プログラム4**のように、指定した引数を文字型・数値型を問わずに欠損値に変えることが出来る call missing ルーチンを使用する。

```
data EX.'集計エリア'n;
 modify EX.'集計エリア'n;
  call missing (of F:);
run;
```

プログラム4:
名前の定義で指定の
セル範囲の値リセット(2)

引数の指定を「of F:」とすることで「F」で始まる変数名が全て対象となり、文字型、数値型を問わずに null となる。

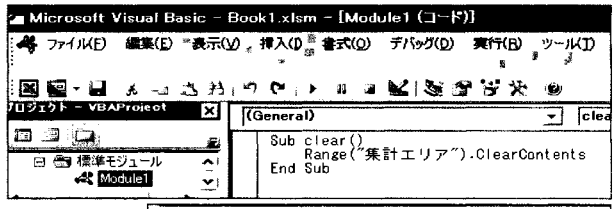
しかし、指定しているセル範囲が大きい場合、値をクリアするだけの単純な処理であっても実行時間がかなりかかってしまう場合がある。このように SAS と EXCEL 間での処理をプログラミングしていると、原因不明の誤動作や、公式の説明にはない仕様の発見により、プログラミングに支障を来すこともある。こうした場合の有力な解決法の一つとして、EXCEL ファイル側に EXCEL VBA で行いたい処理を実行するプログラムをあらかじめ作っておき、SAS の方から必要な場面で呼び出し・実行をする方法がある。

EXCEL VBA は EXCEL を操作することに特化したプログラム言語なので、EXCEL 上の処理に関しては最も柔軟で細かな機能と高速性を有している。

例えば、セルの指定範囲「**集計エリア**」を全て空白化したいのであれば、EXCEL の「開発」タブから「Visual Basic」を立ち上げ、 標準モジュールに以下の EXCEL VBA **プログラム1を**記述する。

```
Sub clear()
    Range("集計エリア").ClearContents
End Sub
```

EXCEL VBA プログラム1:
セルの指定範囲「集計エリア」を全て空白化



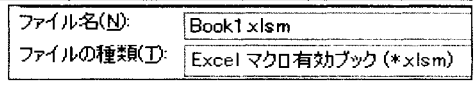


図9 Visual Basic の保存画面

この EXCEL ファイルを保存する際に拡張子を「xlsm」にしておかないと作成した VBA プログラムが保存されないので注意を要する（図９）。

次に、この EXCEL VBA を SAS から実行する**プログラム５**を解説する。

```
libname EX "C:¥Book1.xlsm" header=no scantext=no; ← ①
filename cmds dde 'EXCEL|SYSTEM'; ← ②

data _null_;
 file cmds;
 put'[RUN("clear")]'; ← ③
run;
```

プログラム5:EXCEL VBA を SAS から実行するプログラム

まずファイルの拡張子名が変わったので①でライブラリ名を定義し直しているが、拡張子が xlsx の場合と同様の記述でよい。②で EXCEL の DDE (Dyanamic Data Exchange) という仕組みに FILENAME 文で cmds(何でもよい)というファイル名を割り当てる。DDE は異なるアプリケーション間でやりとりするための仕組みで、③の PUT 文で EXCEL に clear というマクロを実行させることができる。これによって、名前の定義を消さずに、しかも高速にセルのクリアができる。

このように SAS では面倒な、或いは時間がかかる場合は、EXCEL 上の操作や処理を活用して EXCEL

VBA に任せるのも有効な手段である。

VBA で記述する際に、`Sub Auto_Open()` とすれば、EXCEL ファイルを開いた際に自動で実行される VBA となるので、初期設定などの決まった動作であればこの仕組みを利用すると便利である。

なお、EXCEL の「マクロの記録」機能を使えば、ユーザーが手動で行った操作を自動で VBA コードに変換してくれるので、VBA の知識がないユーザーでも簡単に利用することができる。

2.1.4 条件付き書式の設定

SAS と EXCEL を組み合わせてシステムを作る際に、EXCEL 側の機能で、便利なものをいくつか紹介する。

例えば、図 10 の上の図では二つの 3×3 のセル範囲に数字が入力されている。今、左右の表で数字が異なるセルがある場合、図 10 の下の図のように、右側の表の当該セルの文字が自動的に赤い太字になるよう設定したいとする。

「ホーム」タブ→「条件付き書式」→「新しいルール」→「数式を使用して、書式設定するセルを決定」を選び、右側のセル範囲を選択して、式「=RC[-4]<>RC」を設定し（図 11）、条件に合致した場合の書式を選択すると、図 10 の下の図のような結果を得る。

この方法は、Visual S[5] では、数独問題の初期局面で空白だったマス目に新たに見つかった数字を赤色に着色する場合に利用されている。

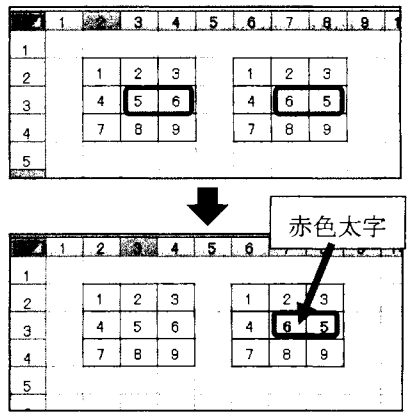


図 10 左右で異なる数字を太字の赤色にする

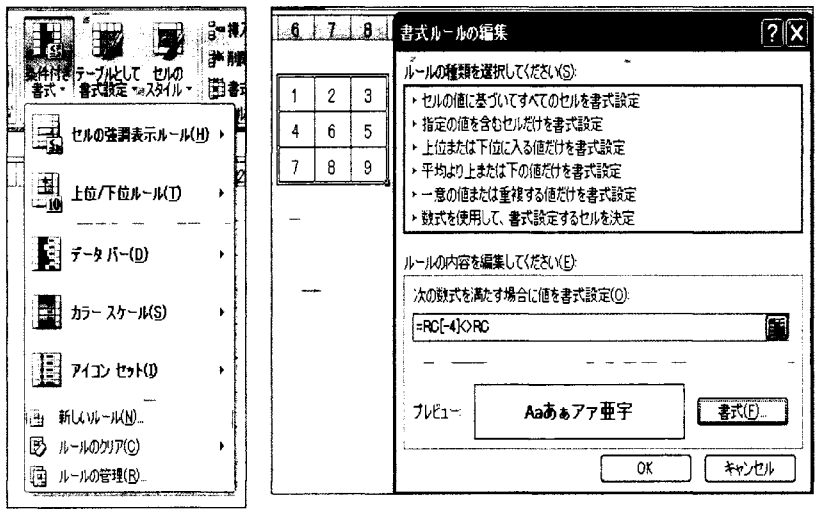


図 11 Excel の操作画面

2.1.5 EXCEL の画面分割

LIBNAME EXCEL で連続して、複数のシートに値を出力する場合、EXCEL ファイルにある複数のシートを同時に画面表示すれば、一画面で確認することができて便利である。Visual S[5] では数独解答中の途中の局面と、フィルタ（3.1 節参照）の変化の様子を同時に可視化するために使用されている。一つの EXCEL ファイルにある複数のシートを同時に画面表示できることは EXCEL のユーザーなら常識であるが、確認の意味で簡単に解説する。

「表示」タブ→「新しいウインドウを開く」（図 12 左）をクリックした後、「整列」で「左右に並べて表示」（図 12 右）をチェックして OK すると、2 つのシートの場合なら、2 つのウインドウに表示される。

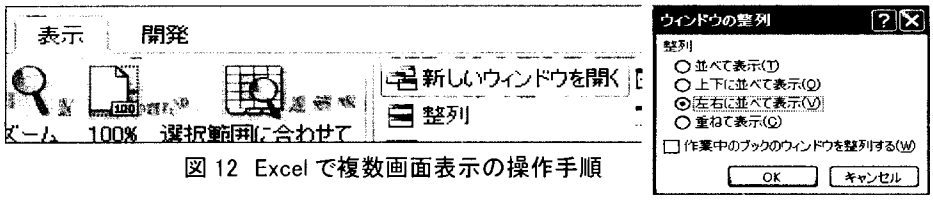


図 12 Excel で複数画面表示の操作手順

図 13 は、同一の EXCEL ファイル内にある「Sheet1」と「Sheet2」を画面分割で表示している。

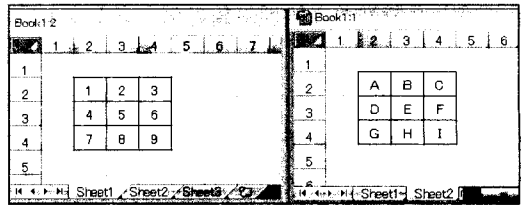


図 13 2つの Excel シートを画面分割表示

2.1.6 クリップボードを経由したデータの出力

LIBNAME EXCEL を使って SAS データセットを EXCEL ファイルに出力する際の留意点の 1 つが、処理時間である。データが数千行や、数百列以上になる場合、LIBNAME EXCEL による出力だとかなり時間がかかってしまう。このような大容量のデータを出力する場合の対処法の一つとしては、LIBNAME EXCEL ではなく、OS のクリップボードを経由させれば、かなり実行時間の短縮を図ることができる。この方法について解説する。

クリップボードとは、文字や画像をコピーした際に、その情報を一時的にストックしておくために OS が利用出来るバッファー領域のことである。ペーストする際はそこからデータが取り出される。

プログラム 6 では「TEST」という名前のデータセットを作成した後、その中身をクリップボードにコピーしている。

①の FILENEME 文でクリップボード(CLIPBRD)に「clip」という任意の名前を付ける。

②で出力先に「clip」、つまりクリップボードを指定して、③で出力する変数名を指定している。

このプログラムを実行すると、クリップボードに「いろは」をコピーしたことになる。従って、EXCEL や WORD はもとより、アプリケーションを問わず、右クリックから「貼り付け」や「paste」を選択すると、「いろは」という文字列が貼り付けられる。クリップボードのデータを EXCEL で指定したセルに張り付ける VBA は、簡単なコードで記述できる。B2 セルにペーストする例を EXCEL VBA プログラム2に示す。EXCEL VBA を作成した後、「2.1.3 EXCEL VBA の利用」で解説した方法を使い、クリップボードにデータを格納した直後に、プログラム 5 (2.1.3 節)に倣って SAS から EXCEL VBA を実行すれば、データセットの内容が EXCEL 画面に表示される。

```
data TEST;
 X='いろは';
run;

filename clip CLIPBRD; ← ①

data _null_;
 file clip; ← ②
 set TEST;
 put X; ← ③
```

プログラム6:
クリップボードに出力する
SAS プログラム

```
Sub paste()
 Range("B2").Select
 ActiveSheet.paste
End Sub
```

EXCEL VBA プログラム2:
クリップボードの情報を
EXCELのセルに
貼り付ける

2.2 SAS データセットの世代管理機能

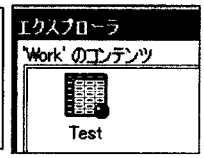
今回我々は数独解法に再帰的プログラミング技法を使用した。2011 年時点のバージョン(略称 SSS)では、数独を解く際に中心的な役割を果たす SAS プログラムの中で、%include 文を使って更に同じプログラムが実行されていた。S[5] 以降のバージョンではそのプログラムがマクロ化されて、そのマクロの中に同じマクロが組み込まれている。SSS では、超難問の解答には実に 1248 回の再帰が発生した(知平・周防 2011、Suoh 2012)。

こうした特性から処理対象のデータセットは、実行中に同じデータセット名で繰り返し上書き・更新されていく。そのような場合、実行後にデータセットがどのように変化していったのかを確認できる仕組みは必要である。なぜならば、もし解けない数独問題が出た場合には、どこまで解答が進み、どの時点で解答作業が行き詰まったかを解明して、システムのバージョンアップをしなければならない。そのために、途中のデータセットをすべて保存するように設計されている。ただし、このデータセットがいつの時点のものであるかは、保存したフォルダの名前を頼りに追跡する必要があり、システム開発者でなければ解析は容易ではない。そこで、今回は数独の解法過程がリアルタイムで画面上に見える様に、進捗状況を EXCEL 画面にその都度表示できるように機能追加を図った。

今回我々が採用した方式には、SAS データセットの「世代管理機能」を利用した。この機能は応用度が高く、非常に便利であるが、日本では知名度が著しく低いので、丁寧に解説する。

図 14 のプログラムを実行すると、SAS のテンポラリーデータセット保存用のライブラリ参照名 WORK にデータセット「TEST」が作成される。

```
data TEST;
 X=2;
run;
```



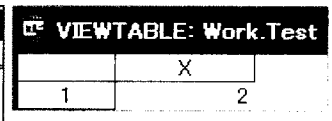
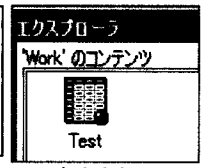


図 14 SAS データセット

続いて、図 15 のプログラムを実行すると、データセットは当然上書き・更新される。この時点で、最初に X=2 であった「TEST」の情報はどこにも残っていない。もしも残したい場合は、データセット名を変えてから実行するしかない。

```
data TEST;
 X=X*X;
run;
```



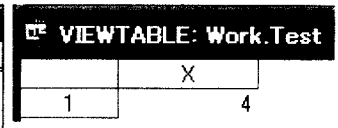
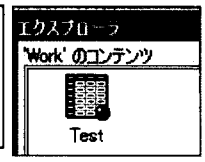


図 15 上書きされた SAS データセッ

では、このデータセット「TEST」を一度削除して、図 16 のプログラムを実行してみる。また同じように「TEST」が作成される。

```
data TEST(genmax=3);
 X=2;
run;
```



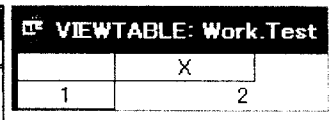
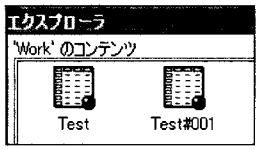
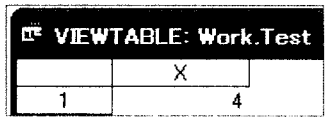


図 16 第一世代 SAS データセット

続いて、図 17 のプログラムを実行する。

```
data TEST;
 X=X*X;
run;
```





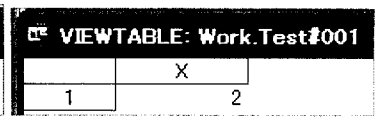


図 17 第二世代 SAS データセット

今度も、X=4 であるデータセット「TEST」が作成されるのは当然だが、その他に、「TEST#001」とい

う名前のデータセットが自動的に作成されていて、中身をみるとX=2となっている。この「TEST#001」が「世代データセット」である。genmaxオプションにより更新前のデータセットが自動的に保存されている様子が分かる。

更にもう一度、図17のプログラムを実行すると、「TEST#002」が作成され、Xの値は8になる。genmaxオプションでは保持する世代の最大数を指定することができる。この例では3に設定しているので（図16)、もしも4回目の更新が行われた場合は、一番古い世代のデータセットが削除され、常に最新の3つの世代のデータセットが保持された状態になる(図18)。

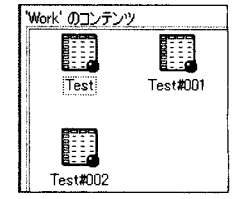


図18 第三世代SASデータセット

```
data A;
 set 'TEST#001'n; ← ①
run;

data B;
 set TEST(gennum=-1); ← ②
run;
```

プログラム7:
世代データセットの名前指定をするSASプログラム

世代データセットはプログラム7の①のように名前で直接指定することもできるが、通常は②のように「gennum=」オプションを使って、何世代前かを指定して使用することが多い。よく使われる例としては、compareプロシジャを使って現在のデータセットと一世代前のデータセット(gennum=-1)を比較して、どこが変更されたかを確認する活用法がある。

3. Visual S[5]の解説と実行方法

3.1 数独の解法アルゴリズム

Visual S[5]の説明の前に、数独の説明と、今回のバージョンアップのベースとなっているS[5]がどのような方法で数独を解いているのかについて、ごく簡単に解説する。

数独のルールは以下の2つである。

(1) 空いているマス目に 1～9 のいずれかの数字を入れる。

(2) 横列(「行」)、縦列(「列」)及び、太線で囲まれた9個の3×3 のボックス(「箱」)内に同じ数字が複数含まれてはいけない。

			8		1			
		6		3		9		
	7						3	
6		7				5		4
			2		6			
2				7				8
			4		5			
3		8			○	4		2
	1		3		9		8	

図19 数独問題例

まず、人間が図19の数独問題を解こうとする際の思考プロセスを考えてみよう。一例を挙げると、同図の○がついているセルに注目する。このセルが属する「列」を見ると、上から順に、1,6,5,9の数字が既に使われている。また「行」を見ると左から順に3,8,4,2が既に使われている。さらに「ボックス」を見ると左隅から時計回りで4,5,9,3が既に使われている。これら既に使われている数字をまとめて整理すると、1から9のうち未だ使われていないのは7しか残っていない。従って、○のマス目の数字は7に確定できる。

このように、人間が数独を解く際に用いる最も普通の思考プロセスは、行・列・箱で既に使用されている数字を全て列挙して、1～9の数字と振るい(フィルタ)にかけて、最後に残った数字を正解の候補とし、もしそれが一個ならその時点で確定できるというものである[1]。マス目が新たに1つでも確定できれば、他のマス目のフィルタ情報に影響を与えることになり、フィルタ情報を更新することができる。この作業を繰り返すことによって一つずつマス目が埋まっていくことになる。

[1] 我々の数独システムでは、全ての空白のマス目に対して、関係する「行・列・箱」に属する確定済みの数字を特定して9桁の数字にまとめた文字変数を「フィルタ」と読んでいる(知平・周防 2011)。

S[5]の興味深い点は、この人間が採る論理的なプロセスと同様の方法を一番目の基本作戦としていることで、人工知能の分野で「ヒューリスティックス」(heuristics)と呼ばれる手法の一種である。

図20では、図19の数独問題例を再度取り上げている。まず与えられた初期局面から「フィルタ」と呼ばれる9桁の数字から成る文字変数を作成する。他のマス目で既に使用されていて、使えない数字は既に0以外の数字が入っている状態となっている。○で囲んだマス目のフィルタの中身は123456089である[2]。これは7以外の数字は既に使われていることを示している。従って、このマス目には7という数字が確定する。

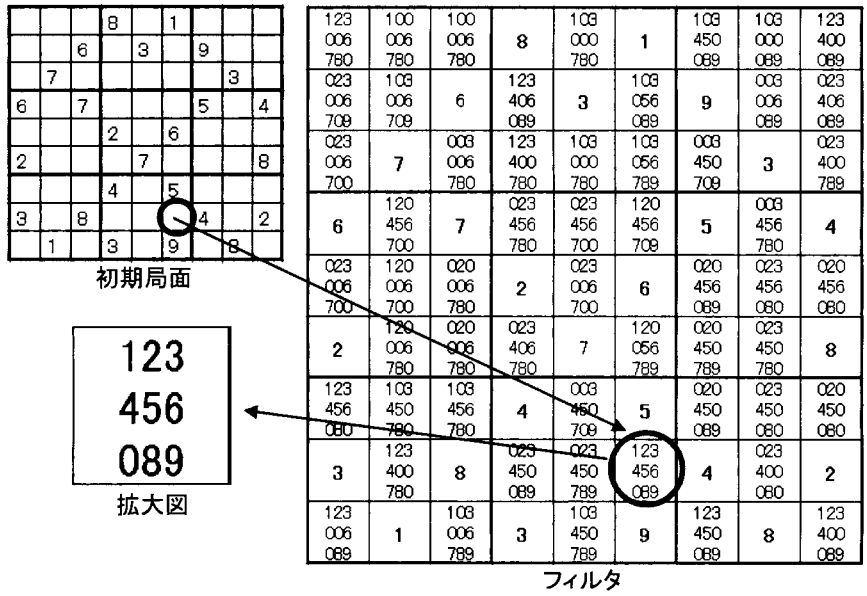


図20 フィルタの更新

今、取り上げたのはS[5]に実装されている6つの解法アルゴリズム(「作戦」)のうちの最初の作戦であり、最も基本的な作戦[3]である。

3.2 Visual S[5]のシステム構成

Visual S[5]については、SAS総会後に全プログラムをネット上に公開する。また、SASユーザー総会の資料ダウンロードページからでもダウンロードできる。実行方法とSASプログラム、及び作成される結果について解説する。ダウンロードした圧縮ファイル(zip形式)を解凍すると、以下のファイルが保存されている。その他に、利用方法などのドキュメントも含まれている。

① VS5.sas・・・・メインとなるSAS数独プログラム
② SUDOKU_SHEET.xlsm・・・・可視化するための出力テンプレート
③ DAT_to_EXCEL.sas・・・・テキストファイルに保存された数独問題をSUDOKU_SHEET.xlsmに展開するためのSASプログラム
④ EXCEL _to_DAT.sas・・・・SUDOKU_SHEET.xlsmで作成した問題をテキストファイルに変換して保存するためのSASプログラム
⑤ フォルダ「problem」・・・・テキストファイル化された数独問題を保存するフォルダ

SASプログラムの実行に先立ち、最初に「SUDOKU_SHEET.xlsm」を開く。セキュリティの警告でマクロの許可を求められる場合は「有効」にする。マクロが有効化されると「Sub Auto_Open()」で定

[2] 図20では見やすさから、9桁の数字を3段に分けて表示している。
[3] 作戦①~作戦⑥については、参考文献(1)、(3)、(5)で詳細に解説している。

義された EXCEL VBA が自動的に実行され、画面が自動的に分割される。これによって複数のシートの内容を同時に見ることができる。開くウインドウの数や大きさはユーザーの好みに応じて調整すればよい。どのように画面を分割しても LIBNAME の動作には影響を及ぼさない。

次に「SUDOKU_SHEET.xlsm」(図 21)の各シートの説明を行う。

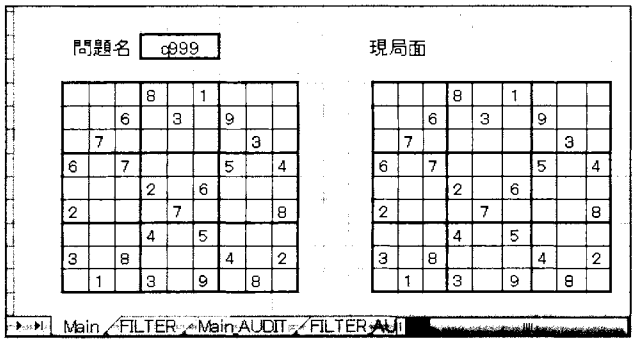


図 21　SUDOKU_SHEET.xlsm (シート Main)

① シート「Main」: 数独の初期局面と解答経過が反映されるメインのシート(図 21)。左側は初期局面で、実行を開始すると右側の「現局面」に新たに数字が埋まっていく。

② シート「FILTER」: フィルタの内容が反映されるシート(図 22)。シート「Main」にある「現局面」を解いている過程でフィルタの中身が常に更新される。

③ シート「Main_Audit」: 数独が完全に解き終わった後に、解答に至るまでの局面変化の履歴が出力されるシートになる。詳しくは後述する。

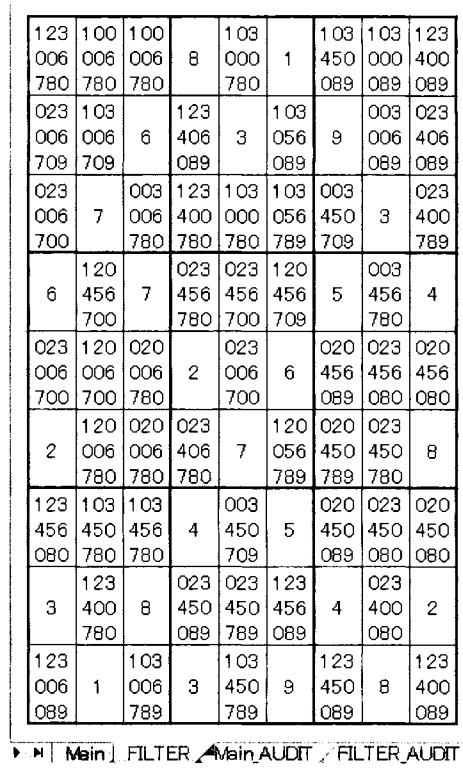


図 22　SUDOKU_SHEET.xlsm (シート FILTER)

④ シート「Filter_Audit」: 数独が完全に解き終わった後に、解答に至るまでのフィルタの変化の履歴が出力されるシートになる。詳しくは後述する。

画面分割はどのようにしても構わないが、プログラム実行中にリアルタイムで更新されるのは「Main」と「FILTER」の 2 シートだけなので、その 2 シートが同時に見えるように分割すれば十分である。

3.3 Visual S^5 の実行プロセス

3.3.1 数独問題の保存と呼び出し

「SUDOKU_SHEET.xlsm」を開いた状態で、「EXCEL_to_DAT.sas」を実行すると EXCEL 上の数独問題を読み込んで、テキストファイルとして保存できる。

「EXCEL_to_DAT.sas」(**プログラム8**) について簡単に解説する。

まず、①で SUDOKU_NAME と指定すると、図 23 の太線で囲ったセル範囲を定義する名前になる。セル結合されているため、問題名(この場合は「q111」)が入力されているのは 4 列目なので②は変数 F4 をマクロ変数 Q に格納する。

③で AREA_1 と指定しているのは太線で囲った範囲の名前になる。9 変数 9 オブザベーションのデ―セットと認識され、変数名 F1～F9 で読み込む。

⑤ で null 値を変換する coalesce 関数を使用して 0 に置き換える。

⑥ で指定した 9 変数を、cats 関数を使用して、空白を含まない 9 桁の数字から成る文字変数に変換している。最後にその値を put して、②で取得した問題名のマクロ変数 Q をファイル名にしてテキストファイル(図 24)を作成している。このようにしてユーザーは EXCEL 上で作成した数独問題を外部ファイルに保存することができる。

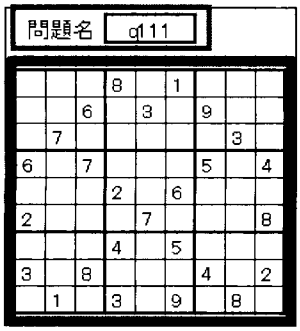


図 23
SUDOKU_SHEET.xlsm

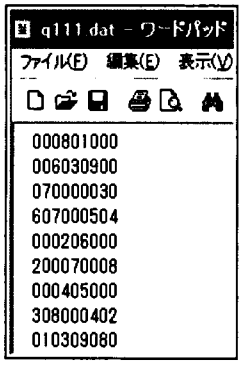


図 24
作成された数独問題のファイル

次に「DAT_to_EXCEL.sas」はそのようにして保存された外部ファイルを EXCEL に展開するプログラムであるが、先の処理の逆をしているだけなので解説は割愛する。

```
/*EXCELを指定*/
libname SUDOKEX "./SUDOKU_SHEET.xlsm"
header=no scantext=no;

/*ファイル名(ファイル名)を取得*/
data _NULL_;
 set SUDOKEX.SUDOKU_NAME; ← ①
 call symputx('Q',F4); ← ②
run;

/*EXCEL上の問題を一旦データセット化*/
/*空白は0に変えておく*/
data temp; keep x;
 set SUDOKEX.AREA_1; ← ③
  array f{9} f1-f9;
  array a{9} a1-a9;
   do i=1 to 9;
    a{i} = coalesce(f{i},0); ← ④
   end;
  x=cats(of a{*}); ← ⑤
run;

/*問題をdatファイルとして保存*/
filename out1 "./problem/&Q..dat";
data _NULL_;
 file out1;
 set temp;
 put x;
run;
```

プログラム8：EXCEL _to_DAT.sas

3.3.2 数独解法プログラムの実行

「SUDOKU_SHEET.xlsm」の「Main」シート（図 21）の左側の EXCEL 表に解きたい数独問題がセットされていれば、後は「VS5.sas」を開いて実行すれば、数独問題を解き始め、図 25 が画面表示される。プログラム中に「options icon」の指定しており、これは実行中に SAS 画面を最小化する効果があるので、EXCEL 画面が見やすくなる。実行が開始されると画面の「現局面」の数字が順次更新され、新しく埋められたマス目は条件付き書式設定の効果によって赤い太字に着色される。画面の右にはフィルタ情報を表示しているが、こちらもフィルタが刻々と更新される様子を見ることができる。図 25 の太線で囲った「現局面」のセルには AREA_2 という名前、点線で囲ったセルには FILTER_VIEW という名前がつけられている。

3.3.3 実行完了と結果の確認

SAS プログラムの最終行に `dm 'post "終了しました"';` と記述している。dm 文はポップアップウインドウで任意のメッセージを表示する命令である。この命令が実行されて「終了しました」というメッセージが表示されると Visual S[5] の処理が全て完了したことがわかる。EXCEL 画面を確認してみると、

全ての数字が埋まり、フィルタの中も全て1桁の数字になっていることがわかる(**付録1**)。

また、シート「Main_AUDIT」や「FILTER_AUDIT」を確認すると、更新履歴が全て縦に連結されて表示され、どのような順番で値が確定されていったかを追跡して確認することができる(**付録2**)。

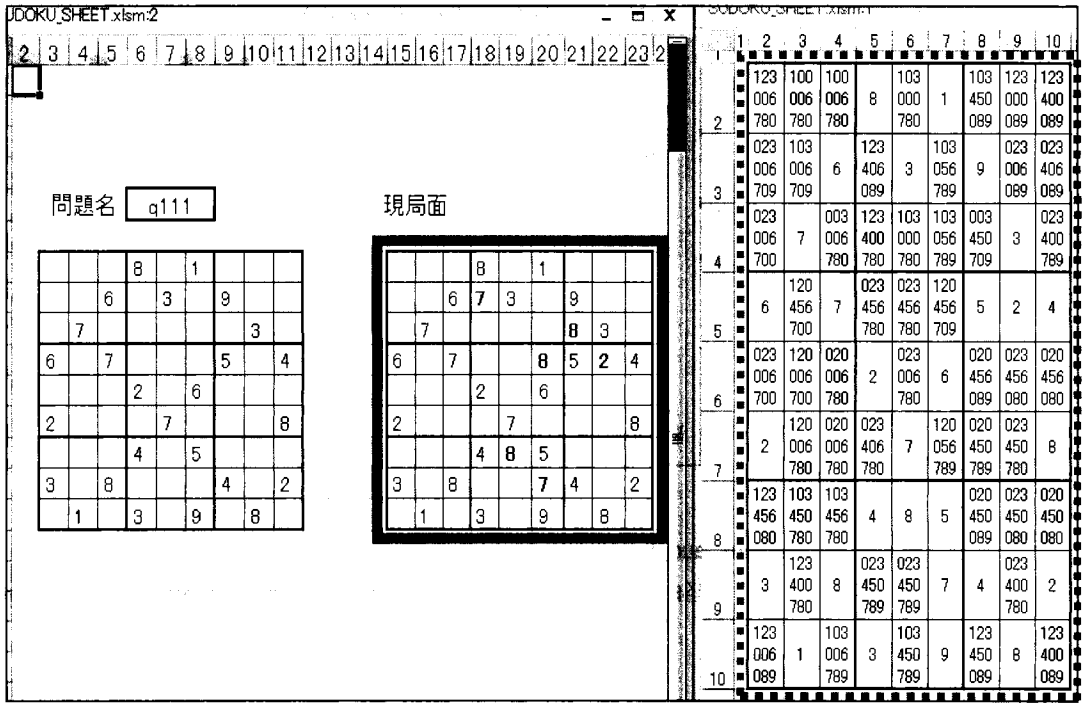


図25　実行中の SUDOKU_SHEET.xlsm (シート Main)

3.3.4 VisualS[5]で追加した機能

Visual S[5]は S[5]に視覚化機能を付加しただけであるが、追加した SAS マクロは4個ある。

一つ目の追加マクロは、解いている途中の局面を EXCEL 画面に反映する SAS マクロ **updateEXCEL(プログラム9)**である。S[5]では再帰処理を行っているため、解答プロセス途中のデータセットは何度も同名で上書き・更新される。これを可視化するには、データセットが上書きされる直前のタイミングで常に EXCEL に出力する処理が働けばいいので、このマクロを、再帰処理が行われる SAS マクロ solve(付録3)の中にある「確定した数字をはめ込む」処理の直後に埋め込んだ。

```
/*EXCELファイルの更新マクロ*/
%macro updateEXCEL;
data SUDOKEX.AREA_2; ← ①
 set original; ← ②
  array aa{9} a1-a9;
  array ff{9} f1-f9;
 modify SUDOKEX.AREA_2;
   do i=1 to 9;
    ff{i}=aa{i};
   end;
run;
%mend updateEXCEL;
```

プログラム9：
SAS マクロ updateEXCEL

①が EXCEL 上の出力範囲、②が S[5]によって更新されるデータセット名であり、データセットの内容をそのままのイメージでエクセルに出力しているだけの単純な処理である。

二つ目の追加マクロは、フィルタの中身を EXCEL 画面 に表示する **updateFILTER** であるが、updateEXCEL とほぼ同じ記述なのでここでは解説を省略する。

三つ目の SAS マクロ **audit_make(プログラム10)は、エクセル**シート Main_AUDIT と FILTER_AUDIT に

更新履歴を縦に連結して出力している。genmax=オプションを上限値1000に設定した上で、更新されるデータセットに追加した。作成される世代データセットには末尾に連番が付与されるため、その規則性を利用して、縦方向に1つのデータセットに連結している。

先頭①のoptions nodsnferr;は存在しないデータセットを指定しても実行エラーとせずに完遂させるオプションである。このおかげで②や③のようにデータセットの存在確認をせずに最大数を網羅的にset文の対象にするような単純な処理が可能となる。

```
options nodsnferr; ← ①
data _null_;
do i=1 to 999;
  filds=cats("'Filter output#".put(i,z3.),"'n");output;
  orids=cats("'original#".put(i,z3.),"'n");output;
 call symputx(cats('f',i),filds);
 call symputx(cats('o',i),orids);
end;
run;

%macro audit_make;
 data filter_audit;
  set
  %do i=1 %to 999; ← ②
   &&f&i
  %end;
 run;

 data origin_audit;
  set
  %do i=1 %to 999; ← ③
   &&o&i
  %end;
 run;
%mend audit_make;
```

プログラム10：SASマクロaudit_make

4つ目の新規SASマクロ**output_audit**は、終了メッセージを出力するコードの直前に追加した。フィルタの更新履歴をクリップボード経由で行うので、高速でエクセルシートFILTER_AUDITに出力する。

3.4 Visual S[5]の今後の課題

探索過程におけるVisual S[5]の更新履歴の確認画面（**付録2**）は、現段階では、更新されたデータセットを作成順に単に縦に連結・表示しただけである。つまり、いわゆるゲーム木の構造を持っていないので、これではプログラムがどのようにして数独を解いているかの経路を辿ることは難しい。このゲーム木の生成状況は大変興味深いテーマである。

本稿の「3.1 数独の解法アルゴリズム」では、候補となる数字が1つに絞れる場合にその数字を確定する過程を説明した。しかし、我々の数独システムには、その方法で候補が1つに絞りきれない場合に、取り得る値の組み合わせから、正解が含まれる可能性がある候補局面を複数個生成し、その候補局面のそれぞれについてゲーム木の探索を進めるという作戦も複数個、実装されている。

我々は、このゲーム木の生成状況をEXCEL画面上に見やすく表示できることを追求しているが、開発に時間がかかり、今回の報告時点までには実現に至っていない。ゲーム木における各更新局面の位置情報は、既に、システム内に保持しており、その情報を基に見やすい可視化を目指して、現在のシステムを更に改良していきたい。

4. おわりに

本論文の当初のメインテーマは、SAS数独プログラムの実行内容をリアルタイムでEXCEL画面に可視化することであった。これができれば、SASの出力用画面として、SAS環境以外のツールが、リアルタイムで利用できることに繋がり、SASの利用環境が格段に拡大することになる。

Visual S[5]はS[5]をベースとして、SASデータセットとして作成される解答途中の数独局面が更新され

る度にEXCEL画面に表示することにより可視化を実現できたが、解答を得るまでの実行時間の点では、数独を解くことに関係のない処理をしている分だけマイナスになり、非合理的である。

では、何のために可視化をするのか。その理由は二つある。第一は、我々の数独プログラムを利用するユーザーが見て素直に楽しめること。第二には、数独問題が解かれていく過程をリアルタイムで見えることで、システムの動きが容易に確認できるので、システム評価が直感的にできることである。

周防は2012年に米国フロリダで開催されたSAS Global Forum 2012で、数独を解く第一世代のSASプログラム「SSS」を発表した。その発表準備期間中にニュージーランド人とカナダ人のグループから、自分たちのグループも数独(Sudoku Puzzle)を解くSASプログラムを作成して発表する(Kastin,M & Tabachneck, A 2012)とのメールがあった。彼らのアプローチは新しいプロシジャFCMPを使って数理的に解く方法であった。解答にかかる実行時間は我々のシステムと比べると格段に早いものであった。当時proc FCMPを知らなかったユーザーとしては大変示唆に富んだ研究報告だった。ただ、もし解答時間の競争なら何もわざわざSASでシステムを組む必要はなく、別の言語で組めばよかったのだが、周防は数独の解答に極めて「不向きな」SASデータセットを使って解きたかったのが、この研究を始めたきっかけである。要するに、SASの強力なデータハンドリング機能を試してみたかったのである。

SASには高度な解析技法が実装されており、それを用いることで極めて豊かな知見を得ることができる。しかし先端の統計解析処理は、時として、各分野の先頭を走る一部のパワーユーザーの間のみでしかその魅力を共有できないという側面がある。

一方でSASのもう一つの大きな柱である強力なデータハンドリング機能については、基礎的な部分であるため、様々な応用ができ、分野を横断して多様なレベルのユーザー同士、共有発展がしやすい。SASを長く使ってきた我々著者としては、次世代のSASユーザーを増やし、そのレベルを上げるために、もっと魅せるSASプログラミングを意識する必要があると考えている。魅せる工夫としては、解析結果のビジュアル化が最近の流行であるが、今回のようなデータステップの途中経過の処理を「可視化」するのも一つの有効な手段ではないかと考える。

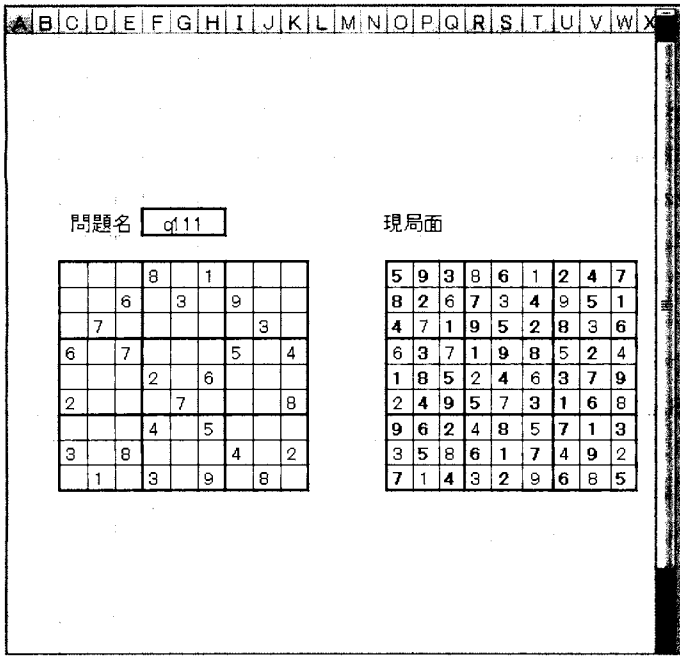
本論文では、こうした問題意識から、SAS歴の浅いユーザーであっても興味を持って読みやすいように、数独解法のアルゴリズムの紹介は最小限にし、視覚化に用いている技法に焦点を当てて解説した。

今回、数独解法過程をある程度可視化できたことで、SASによる魅せるデータハンドリングに一層興味が湧いたので、今後もこうした研究姿勢を続けていくが、その結果、SASプログラミングに更に興味を持つユーザーの増大に繋がれば、著者としてこれほどの喜びはない。

参考文献

(1) 知平菜美子・周防節雄（2011）数独パズルを解くSASプログラム、『SASユーザー総会 2011 論文集』、pp.353-363。
(2) 周防節雄・知平菜美子(2011) ＳＡＳマクロ言語を使った数独パズルを解くプログラムの構造と制御方法、『SASユーザー総会 2011 論文集』、 pp.365-378。
(3) Suoh, Setsuo (2012), Sudoku-Solving System by SAS®, the Digital Proceedings of SAS Global Forum 2012, Florida, USA.
((http://support.sas.com/resources/papers/proceedings12/225-2012.pdf))
(4) Kastin, M. & Tacachneck, A. S. (2012) Yet Another Sudoku Solver: PROC FCMP, the Digital Proceedings of SAS Global Forum 2012, Florida, USA.
(http://support.sas.com/resources/papers/proceedings12/433-2012.pdf)
(5)周防節雄 (2012) SAS®言語で解く数独パズル、『商経学叢』第59巻第2号、近畿大学商経学会、2012年12月、pp.131-167。
(6) 森岡裕（2013）ライブラリ参照と名前定義を利用してEXCELファイルへの柔軟な入出力を実現する方法と応用例の提案—解析結果のレポーティングからセルオートマトンまで—、『SASユーザー総会 2013 論文集』、pp.377-389。

付録1
数独プログラム実行終了時の MAIN シートとフィルタの EXCEL 画面



問題名 q111

			8		1			
		6		3		9		
	7						3	
6		7				5		4
			2		6			
2				7				8
			4		5			
3		8				4		2
	1		3		9		8	

現局面

5	**9**	**3**	8	**6**	1	**2**	**4**	**7**
8	**2**	6	**7**	3	**4**	9	**5**	**1**
4	7	**1**	**9**	**5**	**2**	**8**	3	**6**
6	**3**	7	**1**	**9**	**8**	5	**2**	4
1	**8**	**5**	2	**4**	6	**3**	**7**	**9**
2	**4**	**9**	**5**	7	**3**	**1**	**6**	8
9	**6**	**2**	4	**8**	5	**7**	**1**	**3**
3	**5**	8	**6**	**1**	**7**	4	**9**	2
7	1	**4**	3	2	9	**6**	8	**5**

シート Main

「現局面」は数字が全て埋まり、正解に辿り着いたことを示している。プログラムが解答した数字は EXCEL 画面では赤色の太字で表示されている。

5	9	3	8	6	1	2	4	7
8	2	6	7	3	4	9	5	1
4	7	1	9	5	2	8	3	6
6	3	7	1	9	8	5	2	4
1	8	5	2	4	6	3	7	9
2	4	9	5	7	3	1	6	8
9	6	2	4	8	5	7	1	3
3	5	8	6	1	7	4	9	2
7	1	4	3	2	9	6	8	5

シート FILTER

数独が完全に解けた場合、すべて 1 桁の数字になる。この数字はシート「Main」の「現局面」と同じになる。

付録2　数独プログラム実行終了時の EXCEL 画面

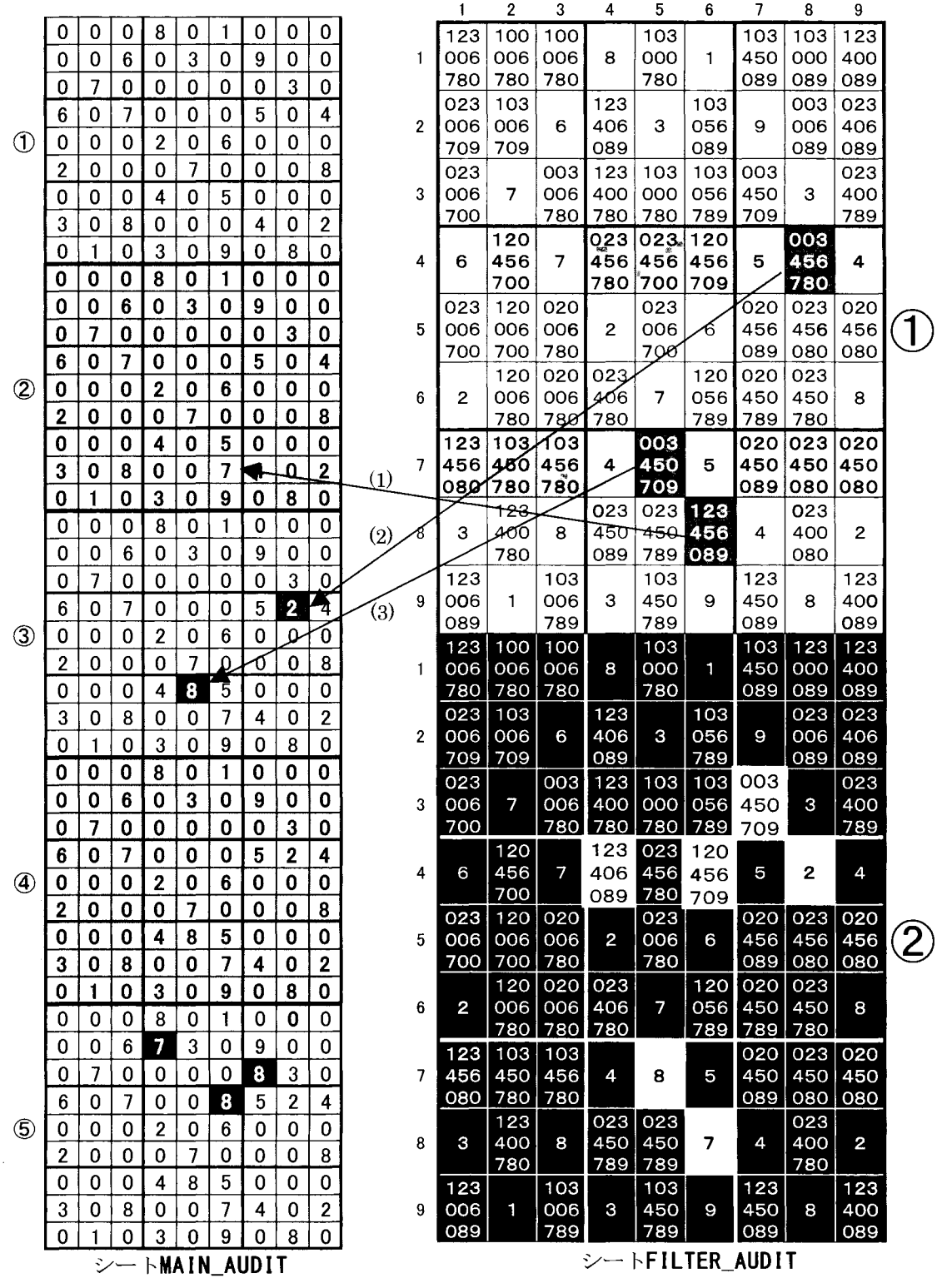
二つのエクセルシート「MAIN_AUDIT」、「FILTER_AUDIT」の中に、正解に向かっている局面とそれに対応するフィルタ情報について、更新記録が逐一保存されている。付録 1 の数独問題をここでも取り上げて、次頁にある解答プロセスの初期段階のこの二つのエクセルシートの更新記録を辿ってみる。

次頁左にあるエクセルシート MAIN_AUDIT には初期局面から順次確定していったマス目が記録されていく。左端の枠外の数字①～⑤は局面[4]の更新段階を示している。右側にあるエクセルシート FILTER_AUDIT は正解途中の各段階でのフィルタ情報を表示している。右端の枠外の数字①と②はフィルタ情報の更新段階を示しており、最初のフィルタ情報は①の 9×9 のセル、次のステップのフィルタ情報は②の 9×9 のセルである。

図の矢印(1)の先に新たに 7 という数字が見つかっているが、これは本文中でも解説した通り対応するフィルタが 123456089 であることから確定できたことがわかる。この数字の確定は、作戦①にある二通りの論理の内の一つ目の適用結果である。

---

[4] 図を見やすくするために、各段階の 81 個のマス目は白地と黒地で交互に色分けしている。右側のフィルタ情報の表示も同様に色分けしている。


①
②
③
④
⑤
(1)
(2)
(3)
1 2 3 4 5 6 7 8 9
シートMAIN_AUDIT
シートFILTER_AUDIT

矢印(2)の先にも新たに２という数字が見つかっているが、これもこのフィルタ情報から確定できた。対応するフィルタ情報は 003456780 なので候補ナンバーは１，２，９の３個あるが、このマス目の「行」を見ると、ほかの空いたマス目のフィルタには全て２が入っているので、それらのマス目には２が使えないことが分かる。従って、この行で２を使うことができるのは矢印(2)の先にあるマス目だけということになり、２が確定する。これは作戦①の二つ目の論理の適用である。矢印(3)の先のマス目も同じ論理で８を確定している。

このようにしてシート「MAIN_AUDIT」のステップ③までで、３つの数字を確定できたが、ステップ④でデータセットが更新されているのにも関わらず、新たに発見された数字はない。これは、この時点まで使用していたフィルタ（「FILTER_AUDIT」ステップ①の情報）の内容からはこれ以上新たに数字を特定できなくなったことを意味する。その時は、この新たに確定した３つの数字を追加してフィルタ情報の更新が行われ、「FILTER_AUDIT」のステップ②となり、次の新しい数字の発見に進んでいく。その結果「MAIN_AUDIT」のステップ⑤で新たに三つのマス目に７，８，８が確定している。これは作戦③～⑥を適宜適用することによって、ゲーム木を探索した結果である。

付録3　新規追加した SAS マクロ

```
%macro solve(row_col_box_no); %* Operation (1);
   data cell;
      keep v row_no col_no box_no found;
      set candidate;
      if zero_cnt NE 1 then return;
      do i=1 to 9;
         if substr(filter,i,1)=0 then do;
            v=i;
            found=1;
         end;
      end;
   run;
   ･･･(中略)･･･
  data original;
        keep a1-a9;
         array a {9} $ 1;
         array new {9} $ 1;
         merge originalx found_nos;
         by mactch_key;

         do i=1 to 9; *新たに確定した数字の保存;
            if new{i} NE . then a{i}=new{i};
         end;
      run;
     %updateEXCEL
   %end;
%mend solve;
```

これまでの S[5] の中に、今回新規に追加した SAS マクロは合計４個である(3.3.4 節)。

左の SAS マクロ **solve** は数独解法エンジンとして本システムの心臓部に相当し、このマクロの中でこのマクロ自身が再帰的に呼び出される。左のコードの中で「新たに確定した数字の保存」をした直後に SAS マクロ **updateEXCEL** を新規に追加した。

なお、フィルタ情報をエクセルシート FILTER に表示する SAS マクロ **updateFILTER** は、行・列・箱のフィルタ情報を統合する SAS マクロ **all_filter** の内部に新規に追加した。

SAS プログラムの全コードは付録４の URL からダウンロードできる。

付録4　数独プログラム Visual S[5] のシステムファイルのダウンロード

本論文で解説した SAS の数独プログラム一式、利用方法、及び関連ドキュメントは、以下の URL からダウンロード出来る。ユーザー登録やパスワードの設定は不要。

http://mighty.gk.u-hyogo.ac.jp/confidential/Visual_Sudoku.zip

Let's Make Forest Plot by SAS

Shinichi Hotta
Development Operations, Pfizer Japan Inc.

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

Abstract:

The tips to create the forest plots easily with SAS/GRAPH's basic functions are introduced.

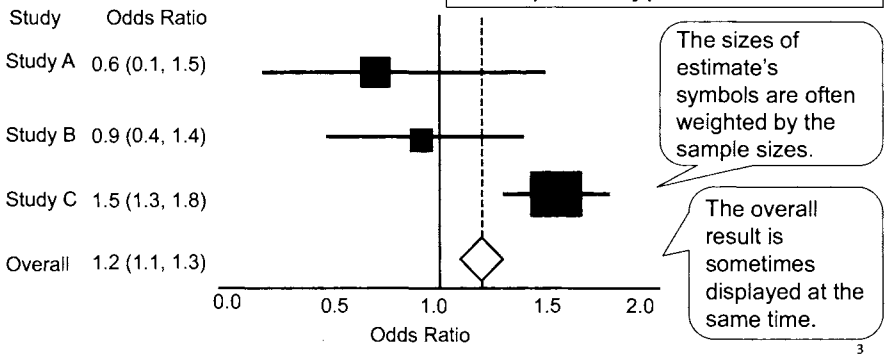
Keywords:

SAS GRAPH, GPLOT Procedure, PLOT Statement, BUBBLE Statement, Annotate Facility

What is "Forest Plot"?

"Forest Plot" is…
- Some groups' estimates and their intervals as a graph. (Anything is OK)
- Used for Meta-analyses.

Example of Typical Forest Plot

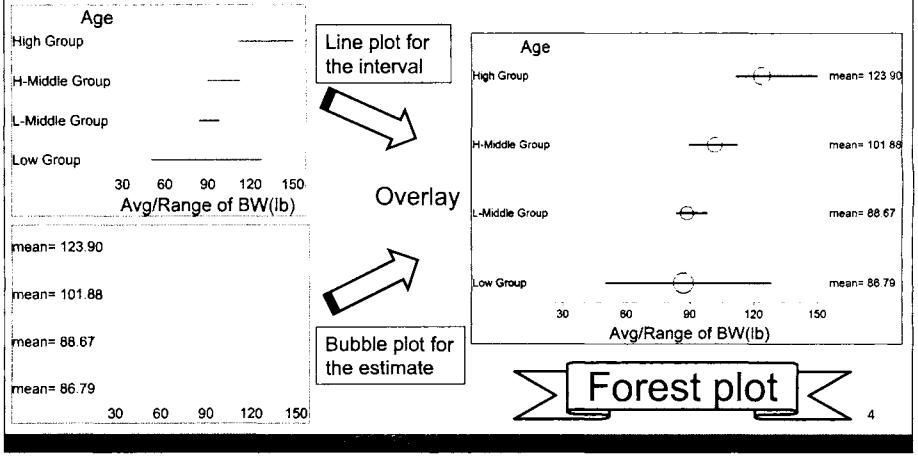


3

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

How to make the forest plot by SAS8.2

- SAS doesn't have the graph function for the forest plot.
- The overlay of Line plot and Bubble plot enables you to make the forest plot.



4

The basic data structure for the forest plot

•The structure of the input dataset is relevant to the forest plot.

(1) The variable to vertically plot left must have the values for the group names (numeric is OK, discrete is required).

(4) At the observations for the interval, the variable for the estimate must be blank.

(3) The variable to horizontally plot must have the values for the estimates and their intervals (upper and lower limits).

	rowid	rowid2	n_subj	bplot
1	H-Middle Group		4	90
2	H-Middle Group		4	112.5
3		H-Middle Group	4	101.875
4	High Group		5	112
5	High Group		5	150
6		High Group	5	123.9
7	L-Middle Group		3	84
8	L-Middle Group		3	98
9		L-Middle Group	3	88.666666667
10	Low Group		7	50.5
11	Low Group		7	128
12		Low Group	7	86.785714286

(5) At the observation for the estimate, the variable for the intervals must be blank.

(Optional)
You can use the aliases for the groups by applying the format to one of the variables to vertically plot.

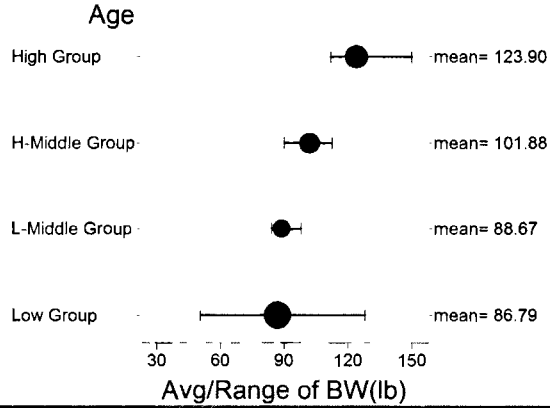
(2) The variable to vertically plot right must have the same values as the same group in the variables to vertically plot left (numeric is OK, discrete is required).

(6) The variable to weight must have the numeric values that mean "N".

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

The forest plot by SAS9.2

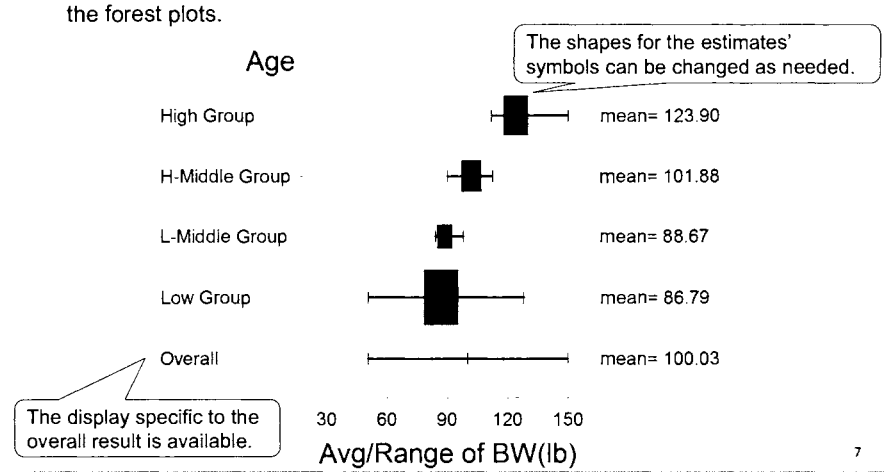
•By SAS9.2 or later version, BFILL= option in GPLOT procedure's BUBBLE statement enables you to fill the bubbles with the colors.
•This looks probably more ideal as the forest plot.



6

The more exploratory techniques

•Instead of BUBBLE Statement, using Annotate Facility will expand the variety of the forest plots.



The expanded basic data structure

•The basic data structure for the forest plot can be expanded and applied into the dataset for Annotate Facility.

The variables in the dataset with the basic data structure are renamed for Annotate Facility.

The added variables for Annotate Facility's settings.

	rowid	yc	size	x	function	text	cbox	xsys	ysys
1	Overall		19	50.5				2	2
2	Overall		19	150				2	2
3		Overall	19	100	symbol	diamond		2	2
4	H-Middle Group		4	90				2	2
5	H-Middle Group		4	113				2	2
6		H-Middle Group	4	102	symbol		black	2	2
7	High Group		5	112				2	2
8	High Group		5	150				2	2
9		High Group	5	124	symbol		black	2	2
10	L-Middle Group		3	84				2	2
11	L-Middle Group		3	98				2	2
12		L-Middle Group	3	88	symbol		black	2	2
13	Low Group		7	50.5				2	2
14	Low Group		7	128				2	2
15		Low Group	7	86.8	symbol		black	2	2

The added observations for the overall results.

The aliases are still applicable through the format .

The highlights of sample SAS programs

Sample for SAS8.2

```
proc gplot data=byage2;
plot rowid*bplot=rowid/vaxis=axis2 haxis=axis1 nolegend;
bubble2 rowid2*bplot=n_subj/vaxis=axis3 bsize=9 bcolor=black;
run;
```

Sample for SAS9.2 or later version

```
proc gplot data=byage2;
plot rowid*bplot=rowid/vaxis=axis2 haxis=axis1 nolegend;
bubble2 rowid2*bplot=n_subj/vaxis=axis3 bsize=9 bcolor=black bfill=solid;
run;
```

Sample for ANNOTATE Facility

```
proc gplot data=anno_age annotate=anno_age;
plot rowid*x=rowid/vaxis=axis2 haxis=axis1 nolegend;
plot2 yc*x=yc/vaxis=axis3 nolegend;
run;
```

If you need the full sample programs, please contact shinichi.hotta@pfizer.com.

医薬品開発、SASシステム

SASハッシュオブジェクトを利用して
医薬品開発に使用するプログラムを効率化する
－有害事象と併用薬、臨床検査値と途中変更のある施設基準値
のマッチングからSASプログラムコードの分析まで

森岡 裕
神田 悟志
ナイフィックス株式会社 バイオメトリクスグループ

Using SAS Hash Objects to Improve Program
-from Matching-Merge Clinical data to Analyzing your SAS Code

Yutaka Morioka
Satoshi Kanda
Biometrics Group, Niphix KK

要旨

SAS9.1 よりデータステップ内でハッシュオブジェクトが利用可能となった。SAS におけるハッシュオブジェクトは、キー(Key)とデータ(Data)から構成され、メモリ上において展開されるルックアップテーブル構造である。

アクセス時間が高速であるメモリを参照することから、主に金融・マーケティング分野のビックデータ関連で効率化のテクニックとして紹介されることが多いようである。

しかし、医薬品開発の分野でも、例えば大規模或いは長期間追跡の臨床試験では臨床検査値データや有害事象の報告数が数百万オブザベーションやそれ以上となることがある。またモンテカルロシミュレーション等では膨大なデータを発生させてから処理を行うケースも多くみられる。

そのような場合、データステップやプロシジャの 1 ステップ 1 ステップが無視できない処理時間を必要とすることが多く、通常のサイズのデータセットに対する処理と同じ感覚でプログラムを書くと、長大な実行時間やディスク容量オーバーなどが原因となり、処理に支障をきたすことがある。

本稿ではそういった問題に対してハッシュオブジェクトの利用を提案する。単純化した例として有害事象データと併用薬データおよび、臨床検査値と施設基準値を効率よくマッチングする方法を紹介する。

また解析業務に使用された SAS のプログラムコードをテキストとして読み込み、Perl 正規表現とハッシュ反復子オブジェクトを使い、どのような処理が行われたかをレビューする方法を紹介し、テキスト処理におけるハッシュオブジェクトの特徴と利用メリットを明らかにしたい。

キーワード：ハッシュオブジェクト　ハッシュ反復子オブジェクト　Perl 正規表現

1. 有害事象と併用薬のマッチング

（1） ハッシュオブジェクト

ハッシュとは連想配列とも訳され、キーとデータから構成される構造を指す。SASにおいて配列という言葉はARRAYステートメントによる配列を示すことが多いが、ハッシュとは大きく性質が異なる。

ARRAYで作成される配列は要素番号によってデータが格納され、要素番号を指定してアクセスされる。それに対して、ハッシュオブジェクトではキーで指定された値をもってデータが格納され、メソッドという特定のコマンドによってのみデータにアクセスされる。

（2） MERGEステートメント、SQLプロシジャとの文法の比較

SASではある変数の値をキーとして複数のデータセットを結合する場合には、いくつかの方法がある。

本稿では有害事象と、それに対して使用された併用薬剤のマッチングをMERGEステートメントおよびSQLプロシジャで行う例を挙げ、同様の処理をハッシュオブジェクトで行う場合について紹介する。

AE(有害事象)

	USUBJID	AEID	AETERM	CMID
1	T-001	1	不眠	1
2	T-001	2	頭痛	2
3	T-001	3	発熱	2
4	T-001	4	胃炎	3
5	T-002	1	腸炎	1
6	T-002	2	感冒	2
7	T-002	3	骨折	3
8	T-002	4	頭痛	3

CM(併用薬)

	USUBJID	CMID	CMTERM
1	T-001	1	睡眠導入薬
2	T-001	2	解熱鎮痛剤
3	T-001	3	胃腸薬
4	T-002	1	胃腸薬
5	T-002	2	感冒薬
6	T-002	3	解熱鎮痛剤

有害事象のデータセット「AE」と併用薬のデータセット「CM」があり、AE.CMIDとCM.CMIDを症例ごとに一致させることによって、各々の有害事象に対して使用された薬剤（いわゆる併用薬）の薬剤名を取得する。

code_1: MERGE

```
proc sort data=AE out=AE_1;
 by USUBJID CMID;
run;

proc sort data=CM out=CM_1;
 by USUBJID CMID;
run;

data MERGE_AE;
 merge AE_1(in=ina)
       CM_1;
  if  ina;
 run;
```

code_2: SQL

```
proc sql noprint;
 create table SQL_AE as
  select AE.USUBJID,AEID,AETERM,AE.CMID,CMTERM
  from AE left outer join CM
  on AE.USUBJID=CM.USUBJID and AE.CMID=CM.CMID;
quit;
```

結果

	USUBJID	AEID	AETERM	CMID	CMTERM
1	T-001	1	不眠	1	睡眠導入薬
2	T-001	2	頭痛	2	解熱鎮痛剤
3	T-001	3	発熱	2	解熱鎮痛剤
4	T-001	4	胃炎	3	胃腸薬
5	T-002	1	腸炎	1	胃腸薬
6	T-002	2	感冒	2	感冒薬
7	T-002	3	骨折	3	解熱鎮痛剤
8	T-002	4	頭痛	3	解熱鎮痛剤

MERGEステートメント、SQLプロシジャ共に、全く同じ結果のデータセットを作成することが可能である。ただし、MERGEステートメントの場合、事前に結合するためのキー変数で、データセットがソートされていることが実行の条件となる。

続いて、ハッシュオブジェクトを使用して、同様の処理を行うコードを提示する。

code_3: HASH(リターンコードなし)

```
data HASH_AE;
 if _N_=0 then set CM;    −①
 if _N_=1 then do;    −②
 declare hash hcm(dataset:'CM');    −③
  hcm.definekey('USUBJID','CMID');    −④
  hcm.definedata('CMTERM');    −⑤
  hcm.definedone();    −⑥
 end;
 set AE;    −⑦
  hcm.find();    −⑧
 run;
```

まず①では、コードの先頭において set ステートメントでデータセット「CM」を指定しているが、_N_=0 の条件を付けているため、オブザベーションは発生しない。これは、変数の定義情報の読み込みを先に済ませておくための処理である。つまり、続けてハッシュオブジェクトに「CM」のデータを格納し、そこからキーで検索して、目的のデータを取り出す処理を行うのだが、ハッシュオブジェクトはデータステップの処理(PDV:プログラムデータベクトル)と独立しているため、先に何らかの方法で変数の初期化を済ませておかないと、データステップの途中で突如、読み込んでいない変数が指定されることによりエラーが発生してしまう。

②について、ハッシュオブジェクトの作成は 1 ステップで 1 度のみ行えばよいため、_N_=1 の条件をつけている。この条件がないとデータセット「AE」のオブザベーションを読み込むたびに、ハッシュオブジェクトを無駄に再作成してしまい、処理時間が遅くなる。

③はハッシュオブジェクトの宣言部分である。declare ステートメントの後に hash と続け、さらに続けて作成するハッシュにつける任意の名前を指定する。ここでは「hcm」という名前でハッシュオブジェクトを定義する。ハッシュオブジェクト名の後の括弧で、格納する変数の抽出先のデータセットを指定できる。データセット名や変数名は全てシングルコーテーションで括る等、通常の SAS データステップと記述法が異なることに注意が必要である。

④から⑥にかけてはメソッドと呼ばれる処理を記述している。ハッシュオブジェクトは通常の SAS の処理とは完全に区別されているため、SAS の通常のステートメントでは一切干渉することができない。数種類のメソッドという決められたコマンド以外では、作成することも、そこからデータを取り出すこともできない。

メソッドに共通の書き方として、まずハッシュオブジェクト名にピリオドを加えて記載して、どのオブジェクトに対してのメソッドであるかを明らかにする。その後にメソッド名を続け、さらに括弧内に引数となる値を記載する。1 メソッドの終わりはセミコロンで表現される。

④は definekey メソッドで、ハッシュオブジェクトのキーを指定する。ハッシュオブジェクトは必ず、キーとデータという構造で定義されるため必須である。なお、後にキーが重複している場合の処理も紹介するが、何も特別な指定をしていない限り、キーが重複していると definekey メソッドはエラーとなる。

例では AE の USUBJID と CMID を使って、CM 内の同一の USUBJID、CMID を探して取得したいため、キーは USUBJID と CMID になる。

⑤は definedata メソッドで、ハッシュオブジェクトのデータを指定する。キーに紐づく内容の部分で、今回は薬剤名を取得したいため、CMTERM を指定している。

⑥は deifinedone メソッドで、これによってハッシュオブジェクトの作成が完了する。

⑦は通常のデータステップで、set ステートメントで AE のデータの読み込みが開始される。

⑧は findメソッドで、これによってハッシュオブジェクトでに設定されたキー(USUBJID、CMID)と、現在の PDV 上の同変数をマッチングして、一致した場合はデータ（CMTERM）に設定されている変数が付与される。

	USUBJID	CMID	CMTERM	AEID	AETERM
1	T-001	1	睡眠導入薬	1	不眠
2	T-001	2	解熱鎮痛剤	2	頭痛
3	T-001	2	解熱鎮痛剤	3	発熱
4	T-001	3	胃腸薬	4	胃炎
5	T-002	1	胃腸薬	1	腸炎
6	T-002	2	感冒薬	2	感冒
7	T-002	3	解熱鎮痛剤	3	骨折
8	T-002	3	解熱鎮痛剤	4	頭痛

先に CM の変数定義情報（ディスクリプタ部）が _N_=0 で読み込みされているため、変数の格納順序が異なるが、結果は MERGE や SQL のものと同一である。

注目点としては、SQL と同様に事前のソートを必要としていない点である。

（3） リターンコード

先述のコード(code_3)でハッシュオブジェクトを使ってマッチングを完了できたが、それは CM に findメソッドで検索される USUBJID と CMID が　全て存在していたからである。例えば CM から CMID=1 のデータを削除し、_CM というデータセットを作成し、それを用いて再度同じコードを実行する。

_CM(CM から CMID=1 を削除)

	USUBJID	CMID	CMTERM
1	T-001	2	解熱鎮痛剤
2	T-001	3	胃腸薬
3	T-002	2	感冒薬
4	T-002	3	解熱鎮痛剤

```
NOTE: データセット WORK._CM から 4 オブザベーションを読み込みました。
ERROR: Key not found.
ERROR: Key not found.
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: データセット WORK.AE から 8 オブザベーションを読み込みました。
WARNING: データセット WORK.HASH_AE は未完成です。このステップは、 8 オブザベーション、5 変数で停止しました。
NOTE: DATA ステートメント 処理（合計処理時間）:
      処理時間                0.32 秒
      CPU 時間                0.04 秒
```

ログに「ERROR:key not found.」というメッセージが現れ、データセットの作成が失敗している。これは AE の T-001 の CMID=1 と T-002 の CMID=1 について、findメソッドでハッシュオブジェクトからデータを取得しようとしたところ、該当するキー情報のデータが存在せず、メソッドが失敗したためである。

このようにたとえキーを見つけられず、メソッドが失敗した場合でも、実行を最後まで完了したい場合はメソッドの先頭に「rc=」というコードをつける。rc は returncode の省略形である。

code_4: HASH(リターンコードあり)

```
data HASH_AE;
 if _N_=0 then set _CM;
 if _N_=1 then do;
 declare hash hcm(dataset:'_CM');
  hcm.definekey('USUBJID','CMID');
  hcm.definedata('CMTERM');
  hcm.definedone();
 end;
 set AE;
  rc=hcm.find();
 run;
```

```
NOTE: データセット WORK._CM 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.AE から 8 オブザベーションを読み込みました。
NOTE: データセット WORK.HASH_AE は 8 オブザベーション、 6 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                0.01 秒
      CPU 時間                0.01 秒
```

	USUBJID	CMID	CMTERM	AEID	AETERM	rc
1	T-001	1		1	不眠	160038
2	T-001	2	解熱鎮痛剤	2	頭痛	0
3	T-001	2	解熱鎮痛剤	3	発熱	0
4	T-001	3	胃腸薬	4	胃炎	0
5	T-002	1	胃腸薬	1	腸炎	160038
6	T-002	2	感冒薬	2	感冒	0
7	T-002	3	解熱鎮痛剤	3	骨折	0
8	T-002	3	解熱鎮痛剤	4	頭痛	0

リターンコードをつけたことで、実行してもログにエラーがでないまま完了することができる。作成されたデータセットをみると、findメソッドが失敗する箇所のrcの値が0以外の数字になっているのが確認できる。

このようにメソッドにリターンコードをつけることで、エラーをrcに数値コードとして受け取り、実行を完了することができる。エラーの内容によってrcに格納される値は変わる。「160038」は「key not found.」に対応する値である。エラーが発生しなければ戻り値は必ず0になる。

そして5オブザベーション目をみると、rcは0以外の値でCMTERMには意図していない値が格納されている。直前の成功したfindメソッドの値を保持して、キーが見つからなかった際に、引き延ばしたいような場合はこのままでよいが、今回の例では誤りとなってしまうので、リターンコードの値を利用してエラー発生の際の処理を付け足す。findメソッドとリターンコードによる処理は、SCL関数を用いた処理、またはインデックスのついた変数とsetステートメント＋key=オプションでマッチングをする際の_IORC_を利用した処理に構造が類似している。

code_5: HASH(リターンコードあり エラー時処理あり)

```
data HASH_AE;
 if _N_=0 then set _CM;
 if _N_=1 then do;
 declare hash hcm(dataset:'_CM');
  hcm.definekey('USUBJID','CMID');
  hcm.definedata('CMTERM');
  hcm.definedone();
 end;
 set AE;
   rc=hcm.find();
   if rc^=0 then CMTRM ='';
 run;
```

	USUBJID	CMID	CMTERM	AEID	AETERM	rc
1	T-001	1		1	不眠	160038
2	T-001	2	解熱鎮痛剤	2	頭痛	0
3	T-001	2	解熱鎮痛剤	3	発熱	0
4	T-001	3	胃腸薬	4	胃炎	0
5	T-002	1		1	腸炎	160038
6	T-002	2	感冒薬	2	感冒	0
7	T-002	3	解熱鎮痛剤	3	骨折	0
8	T-002	3	解熱鎮痛剤	4	頭痛	0

（4） addメソッド

ハッシュオブジェクトにデータを格納する際にデータセットから読み込みだけでなく、メソッドで追加することも可能である。

例として、ある特定の事象名に該当するレコードにフラグをたてる処理を考える。今回は「頭痛」または「骨折」という事象名を対象とする。

前述のcode_3-5ではハッシュオブジェクトに格納するデータがデータセットとして存在していたため、定義情報のセット・変数の初期化の処理を_N_=0の際に実行されるsetステートメントにより行ったが、今回はデータセットから読み込まないため、①で長さを定義し、④で初期化を行うことで同様の処理を実現している。

②でpainという名前のハッシュオブジェクトを作成している。

何のオプションもつけずに作成する場合は空括弧となる。

以下、define.doneまでは前述のコードと同様である。

③及び次の行で、define.doneで作成が完了したハッシュオブジェクトpainに対して、データを格納している。

addメソッドは、括弧内にkeyの場合はkey:'値'、dataの場合はdata:'値'と記述する。

key、data共に複数ある場合は、続けて記述する。

code_6:addメソッド

```
data HASH_PAIN;
 length AETERM $4. FLAG $1.; －①
 if _N_=1 then do;
 declare hash pain();    －②
  rc=pain.definekey('AETERM');
  rc=pain.definedata('FLAG');
  rc=pain.definedone();
  rc=pain.add(key:'頭痛',data:'Y'); －③
  rc=pain.add(key:'骨折',data:'Y');
  call missing(AETERM,FLAG);    －④
 end;
 set AE;
  rc=pain.find();
  if rc^=0 then FLAG='N';
  drop rc;
 run;
```

	AETERM	FLAG	USUBJID	AEID	CMID
1	不眠	N	T-001	1	1
2	頭痛	Y	T-001	2	2
3	発熱	N	T-001	3	2
4	胃炎	N	T-001	4	3
5	腸炎	N	T-002	1	1
6	感冒	N	T-002	2	2
7	骨折	Y	T-002	3	3
8	頭痛	Y	T-002	4	3

2. ハッシュオブジェクトの実行時間

（1）背景

ハッシュオブジェクトを利用する最大のメリットは、プログラム実行時間の短縮である。

データセット同士を結合するために Merge ステートメントを利用する場合には、両データセットからデータを読み込む際に、ハードディスクからメモリにオブザベーションを 1 オブザベーションずつ移動して処理していくため、読み込む量が多い（オブザベーション数が多い、全変数の合計サイズが大きい）とそれだけ処理の完了まで時間がかかる。

それに対してハッシュオブジェクトを利用すると、ステップ実行中はメモリ上にデータを常駐させて処理を行うため、読み込みの時間が著しく速い。ただし、メモリはハードディスクに比べて圧倒的に容量が小さいため、巨大なデータをハッシュオブジェクトとしてメモリに展開するには注意が必要であり、不可能な場合もある。

一般的に、ハッシュオブジェクトが、実行時間短縮に対して効果的だとされているのは、巨大なデータセットに対して小さなデータセットを、共通するキーの値でマッチングするケースである。

本稿では、Merge ステートメント及び SQL、ハッシュオブジェクトによる実行時間を処理対象となるデータセットのサイズを変更しながら比較した。ただし、以降の結果は使用するコンピュータのスペックや実行環境に大きく依存するため、絶対的なものではなく、参考の範囲である。

（2）使用するデータとプログラムについて

code_7:大きいデータセットを作るコード

```
%let size=10;

data BIG;
 do ID=1 to &size;
  do RID=1 to 100;
   array dummy(5) $100.;
        output;
  end;
 end;
run;
```

code_8:小さいデータセットを作るコード

```
data SMALL;
 do RID=100 to 1 by -1;
   array dummy(5) $100.;
        FLAG='Y';
        output;
 end;
run;
```

code_7 で作成されるデータセット「BIG」と code_8 で作成される「SMALL」を RID という共通の変数で結合させる。code_7 の先頭で%let により値を与えているマクロ変数 size の値を増やしていくことで、BIG のオブザベーション数を増やすことができる。また配列で文字変数を作成しているのはデータセットのサイズを増やすためである。上記のデータセットを以下の 3 種類の方法でそれぞれ結合する。

code_9:Merge

```
proc sort data=BIG
          out=BIG_1;
 by RID;
run;
proc sort data=SMALL
          out=SMALL_1;
 by RID;
run;
data MERGE_DS;
 merge BIG_1(in=ina)
       SMALL_1;
 by RID;
 if ina;
run;
```

code_10:SQL

```
proc sql noprint;
 create table SQL_DS as
  select BIG.*,FLAG
   from BIG left outer join SMALL
    on BIG.RID=SMALL.RID;
quit;
```

code_11:HASH

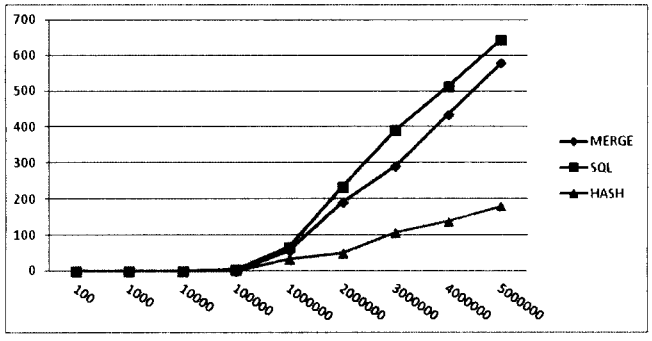
```
data HASH_DS;
 if _N_=0 then set SMALL;
 if _N_=1 then do;
  declare hash HS(dataset:'SMALL');
   rc=HS.definekey('RID');
   rc=HS.definedata('FLAG');
   rc=HS.definedone();
  end;
 set BIG;
  rc=HS.find();
  if rc^=0 then FLAG='';
run;
```

時間の計測については code_9-11 の先頭と末尾で datetime 関数をつかって、日時データをマクロ変数に格納し 2 時点で格納した値の差をもって処理時間(秒)とした。code_9 の Merge についてはソートが処理に必ず必要になることから、sort プロシジャの実行も含めて 1 つの処理とみなした。

(3) 検証結果

方法	データセット「BIG」のOBS数								
	100	1000	10000	100000	1000000	2000000	3000000	4000000	5000000
MERGE	0.02	0.03	0.06	0.44	56.29	190.75	290.6	434.64	578.81
SQL	0.01	0.03	0.08	4.96	66.03	233.86	391.44	513.17	643.74
HASH	0.02	0.02	0.05	0.7	32.29	49.34	107.91	137.33	179.73

(秒)



今回の設定においては、1 万オブザベーションあたりまでは 3 つの処理とも差はほとんどないが、10 万で SQL の処理速度がやや劣った。その後 100 万から 500 万まで計測したところ、Merge と SQL の処理時間の増加率に比べて、ハッシュオブジェクトが明らかに低いことがみてとれた。

500 万オブザベーションにおいて、最も遅い SQL とハッシュの差は 400 秒近くになった。

このことから、ある程度大きいと思われるサイズのデータセットに対して処理を行う場合は、あらかじめダミーデータで処理時間の目安をだしてから、実行時間とコードの可読性等から使用する方法を決定すべきであり、また一定のサイズを超えた場合ハッシュオブジェクトの処理効率がかなり良いことが示唆された。

3. 臨床検査値と途中変更のある検査基準値のマッチング

(1) 背景

多施設共同臨床試験で、各施設において臨床検査値の測定を行っている場合、取得された値を評価するためには、その施設の被験者の性別に対応した基準値を適用する必要がある。また試験期間中に基準値が改定される場合もあるため、検査日と基準値の適応開始日を比較して、各検査日に対応した基準値を取得しなければならない。

マッチング処理のプログラミングとしての難易度はそれほど高くないが、臨床検査値のデータは大規模試験或いは長期間にわたって追跡を行う試験の場合、数百万から場合によってはそれを遥かに超えるオブザベーション数になりえる。データセットのサイズが大きくなると、データステップ、プロシジャステップの 1 つ 1 つにかかる時間が現実的に支障をきたすレベルになる。

ハードを高性能なものにしたり、分散処理を行ったり、オプション等を工夫することで対応できる面もあるが、最も簡単なのは、処理するステップ数を最小限にすることである。

また、各被験者の臨床検査値データは巨大なものになるが、施設基準値のデータセットは如何に多施設共同試験で頻繁に改定されたとしても、比較すれば極小といってよいサイズであるため、ハッシュオブジェクトを利用する上で非常に好条件である。

（2）使用するデータとプログラムについて

本稿で例とするデータセット「LBSS」と「LB」を右に示す。LBSS は施設ごと、項目ごと、また適用開始日ごとにオブザベーションが発生する、例えば SITE「A 施設」の PARAM「RBC」は 2013/1/1 から 2013/01/31 まで LOW(基準値下限)「370」、HI(基準値上限)「480」であり、2013/02/01 からは LOW「375」、HI「485」となる。

例えば、ハッシュを使用せずにマッチングを行う場合 LBSS を SITE、PARAM、を by ステートメントに指定して STDT、UNI、LOW、HI それぞれを transpose プロシジャで転置し、それを元データとマージして、配列で LBDT と

LBSS(各施設の基準値データ)

	SITE	STDT	PARAM	UNI	LOW	HI
1	A施設	2013/01/01	RBC	×10^3/μL	370	480
2	A施設	2013/01/01	WBC	×10^3/μL	3.5	9
3	A施設	2013/02/01	RBC	×10^3/μL	375	485
4	A施設	2013/02/01	WBC	×/μL	3500	9000
5	A施設	2013/03/01	WBC	×/μL	3400	9500
6	B施設	2013/01/01	RBC	×10^3/μL	360	485
7	B施設	2013/01/01	WBC	×10^3/μL	3.6	9.2

LB(各被験者の検査値データ)

	SITE	USUBJID	LBDT	PARAM	AVAL
1	A施設	T-001	2013/01/15	RBC	380
2	A施設	T-001	2013/01/15	WBC	4.2
3	A施設	T-001	2013/02/15	RBC	372
4	A施設	T-001	2013/02/15	WBC	3450
5	A施設	T-001	2013/03/15	RBC	373
6	A施設	T-001	2013/03/15	WBC	3450
7	B施設	T-002	2013/01/15	RBC	500
8	B施設	T-002	2013/01/15	WBC	3.2
9	B施設	T-002	2013/02/15	RBC	380
10	B施設	T-002	2013/02/15	WBC	4.2
11	B施設	T-002	2013/03/15	RBC	371
12	B施設	T-002	2013/03/15	WBC	1.8

code_12:HASH_PATTERN1

```
proc sort data=LBSS;
 by SITE PARAM STDT;
run;

data HASH_PATTERN_1
(rename=(STDT_=STDT UNI_=UNI HI_=HI LOW_=LOW));

 declare hash hlbss(ordered: 'Y'); －①
  hlbss.definekey('SITE','PARAM','COUNTER'); －②
  hlbss.definedata('STDT','UNI','LOW','HI'); －③
  hlbss.definedone();

 do until(ENDLBSS); －④
  set LBSS end=ENDLBSS; －⑤
   by SITE PARAM STDT; －⑥
   if first.PARAM then COUNTER=0; －⑦
    COUNTER+1; －⑧
    rc=hlbss.add();－⑨
 end;

 do until(ENDLB); －⑩
  set LB end=ENDLB;
   COUNTER=1; －⑪
   rc=hlbss.find();－⑫
    do while(rc=0); －⑬
     if LBDT>=STDT then do; －⑭
      STDT_=STDT; －⑮
      UNI_=UNI;
      LOW_=LOW;
      HI_=HI;
     end;
     COUNTER+1; －⑯
     rc=hlbss.find();－⑰
    end;
  output; －⑱
 end;

drop STDT UNI LOW HI COUNTER;
format STDT_ yymmdds10.;
run;
```

code_13:HASH_PATTERN2

```
proc sort data=LBSS;
 by SITE PARAM descending STDT;
run;

data HASH_PATTERN_2;
 if _N_=0 then do;
  set LBSS;
 end;
 if _N_=1 then do;
  declare hash h_lbss(dataset:'LBSS', multidata: 'Y');-①
   h_lbss.definekey('SITE','PARAM'); －②
   h_lbss.definedata('STDT','UNI','LOW','HI');
   h_lbss.definedone();
 end;

 do until(ENDLB);
  set LB end=ENDLB;
   call missing(STDT,UNI,LOW,HI);
   rc=h_lbss.find();

    do while(STDT>LBDT); －③
     rc=h_lbss.find_next();－④
    end;
   output;
 end;

format STDT yymmdds10.;
run;
```

STDT を順番に比較して、LBDT≧STDT{i}の場合に UNI、LOW、HI を書き換えていくような処理が想定される。その場合、LBSS と LB の結合が必要なため、LBSS だけでなく LB に対してのソート処理が必須となる。

それに対して、ハッシュオブジェクトを利用したコードがどのようになるかを 2 種類のパターン(code_12:HASH_PATTERN1、code_13:HASH_PATTERN2)で提示したので以下に詳細を解説する。

まず code_12:HASH_PATTERN1 の処理について説明する。①でハッシュオブジェクト「hlbss」を定義しているが(ordered:'Y')とすることで、作成されるハッシュオブジェクト内部のデータがキーでソートされた状態で格納される。②で SITE PARAM COUNTER と 3 つのキーを定義しているので、この 3 つの値によって③で指定されている変数が昇順にソートされて格納される。ここで COUNTER という LBSS にない変数を指定している点については後述する。

④、⑤の部分について set ステートメントの end=で指定されている変数「ENDLBSS」は LBSS の最終オブザベーションの読み込みで値が「1」となる。それを do until で指定することで LBSS の全オブザベーションが処理されるまでループが発生する。

⑥、⑦、⑧の部分で基準値の改定のあった検査項目については、改定の回数分、変数 COUNTER が+1 される。一度も改定されていない検査項目については COUNTER の値は 1 となる。

そして、⑨で元の LBSS に基準値改定 ID となる COUNTER が追加されたオブザベーションが、add メソッドによりハッシュオブジェクト「hlbss」に追加されていく。

以降⑩から始まる do until ループは LB をセットし、それに対して対応する基準値をマッチングしていく部分になる。⑪でまず COUNTER に初期値 1 を与えて⑫の find メソッドで SITE と PARAM と COUNTER をキーとしてマッチングを行う。今回は、コードの簡素化のため、検査データがあるのに対応する施設基準値がまだ 1 つも登録されていないというケースは除外して想定しているので、ここで必ず初回の施設基準値が取得できる。

⑬の do while ループは rc=0 の間繰り返される。rc=0 とはメソッドが成功している状態である。それは今回のコードでは find メソッドにかかっており、つまり SITE と PARAM と COUNTER でマッチングが成功している限りループするということである。⑰で COUNTER に 1 ずつ加えているので、基準値が改定している数まで漏れなくループが実行されるということである。最終的には+1 により 1 回分過剰に find メソッドを実行しているので最終のメソッドは必ずエラーになるが、その場合正常に find された値が保持され、引き延ばしが起きるため、問題とならない。

⑮で、無事マッチングできた場合に、それを適応するかどうかについて検査日と施設基準値の適応開始日について比較し、検査日>=適応開始日であれば、以下ハッシュオブジェクトのデータ部分を割り当てステートメントで採用する。もし、採用された適応開始日よりもさらに新しい日付で、かつ検査日>=適応開始日を満たすものがあれば、⑰で COUNTER に+1 した後、再度⑱で find メソッドをかけているので、そこで更新される。⑲でマッチングが完了した全オブザベーションが output され、処理が完了となる。

次に code_13:HASH_PATTERN2 の処理について説明する。前提条件として、このコードが正常に実行されるためには SAS9.2 Phase2 以降の環境が必要とする。

まず LBSS を SITE PARAM descending STDT でソートし、各施設各項目の適用開始日が新しい順にデータがくるようにしている。

①でハッシュオブジェクト h_lbss を作成し dataset;で LBSS を指定している。②の部分を先にみると definekey で指定されているのは SITE と PARAM だけとなっており、適用開始日が含まれていないため一意にならない

組み合わせとなっている。そこで①で multidata:'Y'としている。本来ハッシュオブジェクト内のキーは一意にならなければならないが multidata:'Y'とすることで重複を許したハッシュオブジェクトを作成することができる。

SAS 9.3 言語リファレンス:解説編の第 22 章「DATA ステップコンポーネントオブジェクトの使用」P453.非一意キーとデータのペア注釈では、「SAS 9.2 Phase 2 以降では、複数データ項目リスト内の項目は、ユーザーが各項目を挿入した順番で維持されます。」と記載されているため、本コードにおいては直前のソート順にデータが格納されているはずである。

以下は code_12 から COUNTER 変数の処理を抜いたような形になっているが④で find_next メソッドを使用しているのが特徴で、これはハッシュオブジェクト内の次の値を参照するメソッドである③のループ条件と合わせることにより、code_12 とは逆に、最新の適用開始日から遡って値を更新する処理となっている。code_12、code_13 の結果は、変数の並び順については制御するステートメントを入れていないため並びが違うが、内容は同じとなる。

code_12 の結果

	SITE	PARAM	rc	USUBJID	LBDT	AVAL	STDT	UNI	LOW	HI
1	A施設	RBC	160038	T-001	2013/01/15	380	2013/01/01	×10^3/μL	370	480
2	A施設	WBC	160038	T-001	2013/01/15	4.2	2013/01/01	×10^3/μL	3.5	9
3	A施設	RBC	160038	T-001	2013/02/15	372	2013/02/01	×10^3/μL	375	485
4	A施設	WBC	160038	T-001	2013/02/15	3450	2013/02/01	×/μL	3500	9000
5	A施設	RBC	160038	T-001	2013/03/15	379	2013/02/01	×10^3/μL	375	485
6	A施設	WBC	160038	T-001	2013/03/15	3450	2013/03/01	×/μL	3400	9500
7	B施設	RBC	160038	T-002	2013/01/15	500	2013/01/01	×10^3/μL	360	485
8	B施設	WBC	160038	T-002	2013/01/15	3.2	2013/01/01	×10^3/μL	3.6	9.2
9	B施設	RBC	160038	T-002	2013/02/15	380	2013/01/01	×10^3/μL	360	485
10	B施設	WBC	160038	T-002	2013/02/15	4.2	2013/01/01	×10^3/μL	3.6	9.2
11	B施設	RBC	160038	T-002	2013/03/15	371	2013/01/01	×10^3/μL	360	485
12	B施設	WBC	160038	T-002	2013/03/15	1.8	2013/01/01	×10^3/μL	3.6	9.2

code_13 の結果

	SITE	STDT	PARAM	UNI	LOW	HI	USUBJID	LBDT	AVAL	rc
1	A施設	2013/01/01	RBC	×10^3/μL	370	480	T-001	2013/01/15	380	0
2	A施設	2013/01/01	WBC	×10^3/μL	3.5	9	T-001	2013/01/15	4.2	0
3	A施設	2013/02/01	RBC	×10^3/μL	375	485	T-001	2013/02/15	372	0
4	A施設	2013/02/01	WBC	×/μL	3500	9000	T-001	2013/02/15	3450	0
5	A施設	2013/02/01	RBC	×10^3/μL	375	485	T-001	2013/03/15	379	0
6	A施設	2013/03/01	WBC	×/μL	3400	9500	T-001	2013/03/15	3450	0
7	B施設	2013/01/01	RBC	×10^3/μL	360	485	T-002	2013/01/15	500	0
8	B施設	2013/01/01	WBC	×10^3/μL	3.6	9.2	T-002	2013/01/15	3.2	0
9	B施設	2013/01/01	RBC	×10^3/μL	360	485	T-002	2013/02/15	380	0
10	B施設	2013/01/01	WBC	×10^3/μL	3.6	9.2	T-002	2013/02/15	4.2	0
11	B施設	2013/01/01	RBC	×10^3/μL	360	485	T-002	2013/03/15	371	0
12	B施設	2013/01/01	WBC	×10^3/μL	3.6	9.2	T-002	2013/03/15	1.8	0

4. ハッシュ反復子オブジェクトでコードをテキストマイニングする

（1）背景

巨大なデータセットと小さいデータセットをマッチングするのに効率がいいハッシュの性質は、長大なテキストデータから、特定のパターンを抽出して何らかの解析を行うような、いわゆるテキストマイニング処理に適しているため、利用されることが多い。

臨床試験の統計解析結果は通常、1 つの解析出力結果ごとに、それを作成する 1 つの SAS プログラムがあり、その仕様は解析プログラム仕様書等のドキュメントで規定されている。

1 つ 1 つの SAS プログラムファイルの中でどういったプロシジャが使われ、どういった関数が使用されているかを抽出し、それがプログラム仕様に準じているかを簡単にチェックすることができるとすれば、解析プログラムの品質をチェックする上で有用だといえる。

そこで今回は特定のフォルダの中にある複数の SAS プログラムファイルから、プロシジャまたは SAS 関数が使用されている箇所を抜き出し、集計するプログラムについて紹介する。

（2）使用するデータとプログラムについて

右図のように一つのフォルダにまとまって解析に仕様される SAS プログラムファイルが保存されているとする。

code_14:テキスト読み込み部分

```
%let place =C:¥解析 Program;
filename FOLDER pipe "dir / b &place."; －①
data FILELIST;
  length FNAME $200 FPATH $1000;
  infile FOLDER dlm = '09'x;
  input FNAME;
  if index(FNAME,".sas") > 0;
   FPATH = trim(left("&place.")) || "\" || trim(left(FNAME));
run;
filename INP dummy; －②

③
data CODE;
  length LINE $1000.;
  set FILELIST;
  infile INP
   filevar=FPATH end=FLG dlm='09'X missover lrecl = 1000;
  do until(FLG);
    input LINE $;
     output;
  end;
run;
```

フォルダ「C:¥解析 Program」の中身

FIGURE_1.sas	FIGURE_2.sas
LIST_1.sas	LIST_2.sas
LIST_3.sas	LIST_4.sas
LIST_5.sas	LIST_6.sas
LIST_7.sas	LIST_8.sas
TABLE_1.sas	TABLE_2.sas
TABLE_3.sas	TABLE_4.sas
TABLE_5.sas	TABLE_6.sas
TABLE_7.sas	TABLE_8.sas

データセット「CODE」の一部分

	LINE	FNAME
861	call symputx('obs',_N_);	LIST_4.sas
862	run;	LIST_4.sas
863		LIST_4.sas
864	data OUTPUT_1;	LIST_4.sas
865	merge OUTPUT_0(in=ina)	LIST_4.sas
866	ADSL(keep=SUBJID REGAGE SEX);	LIST_4.sas
867	by SUBJID;	LIST_4.sas
868	if ina;	LIST_4.sas
869	run;	LIST_4.sas
870	proc sort;	LIST_4.sas
871	by SUBJID VISITNO;	LIST_4.sas
872	run;	LIST_4.sas
873	data OUTPUT;	LIST_4.sas
874	set OUTPUT_1;	LIST_4.sas
875	by SUBJID VISITNO;	LIST_4.sas
876	if first.SUBJID then do;	LIST_4.sas
877	SEX_=ksscan(put(SEX,HSEX.),2,'');	LIST_4.sas
878	AGE_=REGAGE;	LIST_4.sas
879	end;	LIST_4.sas
880	run;	LIST_4.sas
881		LIST_4.sas
882	/*==========	LIST_4.sas

①から②までで指定のフォルダ内の拡張子が「.sas」のファイル名とパスを取得して、③以降で取得した全てのファイルについて 1 行を 1 オブザベーションとして読み込んでいる。結果は右上のデータセットのようになる。

そこで抽出されたデータセットから関数とプロシジャが使用されている部分を抽出するのが右の code_15 のプログラムとなる。

まず関数をどのように抽出するかについて考える。①から③の部分でディクショナリテーブルビューのうち「VFUNC」を指定して、関数名を取得している。

VFUNC にはその SAS 環境で使用可能な関数の情報が全て格納されている。

あとは取得した関数リストとデータをどうやって照合するかである。

本稿ではパターンマッチングの方法として関数、プロシジャともに Perl 正規表現を使用している。(⑩⑪⑬⑭⑮)

Perl 正規表現はテキストのパターンを規則に従って記号化し、それを SAS の正規表現用の関数に与えてテキストの検索や

code_15:ハッシュ反復子オブジェクトと正規表現による抽出

```
proc sort data=SASHELP.VFUNC－①
          out=FUNC_DIC(keep=FNCNAME) nodupkey; －②
      where FNCNAME^=''; by FNCNAME;
run; －③

data FUNCTION(keep=FNCNAME LINE FNAME)
      PROCEDUR(keep=PRONAME LINE FNAME);
 if _N_=0 then do;
  set FUNC_DIC;
 end;
 declare hash FHUSH(dataset:'FUNC_DIC',ordered:'Y'); －④
  declare hiter FHITER('FHUSH');    －⑤
   FHUSH.definekey('FNCNAME');    －⑥
   FHUSH.definedata('FNCNAME');    －⑦
   FHUSH.definedone();
 do until(ENDLINE);
  set CODE end=ENDLINE;
   rc=FHITER.first();  －⑧
    do until(rc); －⑨
     prx1=compress('m/¥W'||lowcase(FNCNAME)||'¥(/');    －⑩
     if prxmatch(prx1,lowcase(LINE))>0 then output FUNCTION; ⑪
     rc=FHITER.next();  －⑫
    end;
   prx2=prxparse("s/proc (.+?)( |;).*/$1/");  －⑬
   pos=prxmatch(prx2,lowcase(LINE));    －⑭
   if pos>0 then do;
    PRONAME=prxposn(prx2,1,lowcase(LINE));    －⑮
     output PROCEDUR;
    end;
  end;
stop;
run;
```

置換を行うものである。

SAS における Perl 正規表現利用の詳細については SAS のヘルプ等を参照していただきたい。ただし、1 点正規表現で大文字小文字どちらでも検索にかける/i を使用せずに lowcase 関数を使用しているのは、速度の問題として/i パターンよりも相当に処理が速いためである。

さて、関数情報のみを④でハッシュオブジェクト FHUSH に格納し、プロシジャ情報についてはハッシュオブジェクトを作成していないのは、関数は入れ子構造をとることができ、1 つの行に複数回出現する可能性があるため、1 行ごとに全関数リストとマッチングする必要があるが、プロシジャは入れ子構造をとらず、一般的なコーディング規則において 1 行に複数回出現することはないからである。

そのためプロシジャはハッシュオブジェクトとは関係せずに通常のループの中で⑬の正規表現で、proc の後に出現するワードとして抽出してデータセット「PROCEDUR」に出力している。

ハッシュオブジェクト FHUSH に格納されている関数リストとデータをマッチングする上で、通常の Find メソッドは使用することができない。マッチングするキー情報以外の文字が混合しているためである。そのためハッシュオブジェクトのままでは今回の処理にはうまく適用できない。

そこでハッシュオブジェクトをさらにハッシュ反復子オブジェクトという構造の中に格納する。それが⑤の部分である。declare hiter ハッシュ反復子オブジェクト名('ハッシュオブジェクト名');で作成することができる。ハッシュ反復子オブジェクトは、ハッシュオブジェクトを内包した構造体で、通常の key-data に依存しない処理を行うことができる。

具体的には⑧の first メソッドと⑫の next メソッドがあげられる。これはキーマッチングではなく、単純にハッシュ反復子オブジェクト内の先頭をみる処理と、その次をみていく処理である。ここでのリターンコードは単純にオブジェクト内に参照データが存在するかどうかであるため、⑨のループ終了条件は、オブジェクト内の全てのデータを参照することと同義になる。すなわち、あたかも ARRAY 配列を横に順次参照する処理を縦に置き換えたような状態になる。

そして⑩でハッシュ反復子内のデータ、すなわち関数リストに括弧'('を加えた正規表現パターンでマッチしたら OUTPUT するとしたことによって、関数がどれだけ入れ子になっていても全てデータとして起こすことが可能となる。

抽出されたプロシジャ名と使用コード行

	LINE	FNAME	PRONAME
1	proc lifetest;	FIGURE_1.sas	lifetest
2	proc sort data=BASE_0;	FIGURE_1.sas	sort
3	proc sql noprint;	FIGURE_1.sas	sql
4	proc sort data=base_2by axis_time pfs_mrun;	FIGURE_1.sas	sort
5	proc sort data=layout;	FIGURE_1.sas	sort
6	proc sort data=base4_1;	FIGURE_1.sas	sort
7	proc template;	FIGURE_1.sas	template
8	proc sgrender data=base template=FE002;	FIGURE_1.sas	sgrender
9	proc lifetest;	FIGURE_2.sas	lifetest
10	proc sort data=BASE_0;	FIGURE_2.sas	sort
11	proc sql noprint;	FIGURE_2.sas	sql
12	proc sort data=base_2by axis_time OS_mrun;	FIGURE_2.sas	sort
13	proc sort data=layout;	FIGURE_2.sas	sort
14	proc sort data=base4_1;	FIGURE_2.sas	sort
15	proc template;	FIGURE_2.sas	template
16	proc sgrender data=base template=FE001;	FIGURE_2.sas	sgrender
17	proc sort data=CRF.PERFOM out=PERFOM;	LIST_1.sas	sort
18	proc sort data=CRF.PS out=PS;	LIST_1.sas	sort
19	proc sort;	LIST_1.sas	sort
20	proc sort data=LAB;	LIST_2.sas	sort
	proc transpose data=LAB		

抽出された関数名と使用コード行

32	KSCAN	UNIT=kscan(PARAM,2,'()');	LIST_2.sas
33	KSCAN	PARAM=kscan(PARAM,1,'(');	LIST_2.sas
34	IN	where SUBJID='RCC-ZIG04' and COL_11 not in('2011/05/04','2011/07/01');	LIST_2.sas
35	COMPRESS	if IRGDTC^='' then COL_11=compress(COL_11||'('||IRG	LIST_2.sas
36	SYMPUTX	call symputx('obs'_N,);	LIST_2.sas
37	KSCAN	JIKI=kscan(put(VISITNO,GNV.),2,'');	LIST_2.sas
38	INDEX	if index(PARAM,'検査日')>0 then do;	LIST_2.sas
39	KSCAN	VISITNO_=kscan(put(VISITNO,GNV.	LIST_3.sas
40	KSCAN	SJIKI_=kscan(put(SJIKI,GNM.),2,'');	LIST_3.sas
41	KSCAN	KOUHAN_=kscan(put(KOUHAN,GN	LIST_3.sas
42	ROUND	if DT^='' then DT_=put(round(DT,0.01),8.2);	LIST_3.sas
43	SYMPUTX	call symputx('obs'_N,);	LIST_3.sas
44	KSCAN	VISITNO_=kscan(put(VISITNO,GNV.	LIST_4.sas
45	KSCAN	ING_=kscan(put(ING,GNF.),2,'');	LIST_4.sas
46	SYMPUTX	call symputx('obs'_N,);	LIST_4.sas
47	KSCAN	SEX_=kscan(put(SEX,HSEX.),2,'');	LIST_4.sas
48	KSCAN	VISITNO_=kscan(put(VISITNO,GNV.	LIST_5.sas
49	SYMPUTX	call symputx('obs'_N,);	LIST_5.sas
50	COMPRESS	/* if i in(1,2,5) then	LIST_6.sas

抽出終了後は、たとえばプロシジャ名や関数名で集計することで、どういったものを多く使用しているかなどの分析も可能となる。

5. おわりに

本稿ではハッシュオブジェクトの基本的な文法と、応用法を紹介することで今後、医薬品開発の分野でハッシュオブジェクトの利用が進むことを目的とした。今後、グローバルな開発による大規模臨床試験の実施数増加や個別化医療のためにシミュレーションベースの探索的解析の普及、または副作用データベース等を活用したリスク検出などを背景として、医薬品分野で扱うデータの量が劇的に増加していくことが考えられる。

そういった中で、SAS をどのように使っていくのかについて、選択できる方法をより多く知っていることは大きなメリットをもたらすと思われる。

ハッシュオブジェクトの文法、処理のアルゴリズムは SAS と大きく異なる。しかし、異なるが故に組み合わせることで、従来では考えられなかった柔軟、或いは高速な処理が可能となる。

今後もハッシュオブジェクトの利用について継続的に研究していきたい。

6. 文献

1) Art Carpenter(2012). Carpenter's Guide to Innovative SAS Techniques, SAS Institute

2) SAS 9.3 言語リファレンス:解説編

3) SAS 9.3 コンポーネントオブジェクト: リファレンス

SASユーザー会活動の紹介

SASユーザー総会論文集の無料一般公開のインパクト

高橋 行雄
BioStat 研究所（株）

Impact of Open to the Public for Free of SAS User General Meeting Collected Papers

Yukio Takahashi
BioStat Research Co.,Ltd.

要旨　多くの学会紙に掲載された論文が，独立行政法人科学技術振興機構（JST）が提供する「科学技術情報発信・流通総合システム」（J-STAGE）を通じて無料でダウンローできるようになりつつある．SASユーザー総会の論文集には査読付きではないが多くの貴重な論文が掲載されている．これらの論文を引用したいと思っても論文集の現物が手元にない限り困難である． SAS プレミアムラウンジから発表時のスライドなどが，ダウンロードできるようになってはいるが断片的である．そこで，2013 年の SAS ユーザー会の世話人会で論文集の電子的な公開を提案したところ，J-STAGE 活用したらどうかとの提言もあったが，利用資格の条件に該当しないこともあり，自前で対応することが了承された．SAS ユーザー総会は 1982 年に始まり 2013 年で 32 回目であり，現物の収集については，世話人会のメンバーにお願いすることになった．公開に際し，作業量を勘案して論文ごとの対応ではなく 1 冊の論文集まるごとの対応とした．試験的に 1982 年～1989 年までの 8 年分を SAS ユーザー会のトップページに Excel 化した著者索引とともに掲載したところ，外部検索エンジンから検索でき，該当する論文集の PDF もダウンロードできることも確認できた．さらに，PDF の品質を向上しつつ容量を削減するための試行を 2000 年から 2003 年の 4 年分について行った．これらの経験を踏まえ， OCR による自動テキストの付与も加えても PDF に最適化を施すことにより，2013 年の 567 ページの論文集で 16M バイト，1 ページ当たり 30K バイトに圧縮が実現できた．公開された Excel の目次および著者索引を元に，様々な SAS 論文集の活用法を紹介する．

キーワード:　SAS ユーザー総会，一般公開，SAS 論文集

1. はじめに

昨年 2013 年のSASユーザー総会のプログラム編成が 2013 年 6 月 18 日 WEB 公開された後，世話人会にSASユーザー総会論文集の電子公開を提案した．これは，ユーザー会事務局との次のような対話があったことによる．

Q1．高橋：　SAS 関連の高橋これまでの論文（PDF）はどこからダウンロードできないがどうなっているのかるのか．外部検索エンジンでも全く検索できない．

A1．事務局：プレミアムラウンジからダウンロードできると，担当が言っていました．

高橋のアクション：2012 年「SAS プレミアム」で検索するとプレゼン PPT にたどり着くことができた． 2010 年のもようやく別途検索をし，当日の PPT にようやくたどり着けた．ところで，論文はどこだ！_発見できない．

2009年のも発見できない. もちろん以前のも. SASユーザー総会のホームページに, 有料の複写サービスで論文を手に入れることが可能とのアナウンスが次のように掲示されている.

> SAS ユーザー総会 論文集の在庫販売は終了いたしました。
> 1997 年以降の論文については、こちらより有償にて入手する事が可能です。
> 独立行政法人　科学技術振興機構　情報資料館　複写センター
> TEL 0120-004-381　　FAX 03-3979-2210
> http://pr.jst.go.jp/outline/location.html
> なお、新刊につきましては、毎年のユーザー総会開催の際に購入申し込みを受け付けます。申込み数のみの印刷となりますので、ご了承ください。

そこで, JSTのWEB上で「SASユーザ」で検査すると996件の文献が登録されていることか確認された. 著者名による絞込みも 図 1 に示すようにできるようになっていることが確認された. ただし, 内容を確認することはWEB上ではできないので, 1論文あたり約1,000円の複写サービスで現物を入手する必要がある. 筆者の文献数は, 31 件となっていて, 11 件分足りないが網羅的に収集され複写サービスが受けられるようになっている.

図 1　(独)科学技術振興機構のWEB上にある「SASユーザ」関連文献

特定のテーマで文献を網羅的に収集しなければならなければ, 有料の複写サービスであっても利用できるようになっていることはうれしいことではある. しかしながら, 多くの学術雑誌の文献が無料で即時ダウンロードできるようになりつつある時代にあって, 昔ながらの複写サービスとは, さみしい限りである. J-STAGE に収録されている多くの学会の論文は, WEB の検索エンジンで検索可能となっているが, JST に登録されている SAS 関連の文献タイトルは, WEB の検索エンジンでは参照されないことも不満足である.

2. 論文集の電子化と公開

SASユーザー総会論文の電子化は, すべて「自炊」によって行った. 筆者の手元に 32 年分のうち 22 年分があった. 残りの 11 年分は世話人会のメンバー, 筆者の知人などの協力をえて総て収集することができた. そのうち 2007 年度は, 論文集としてではなく, 個別の論文が電子的 PDF で提供されたので, 論文集としての体裁がなかったので, 「論文集」としての体裁に整えることにした. 電子化に際し, ドキュメントの品質とサイズのトレードオフを考慮し, 事前に最適化をはかる必要があったが, ともかく試行することにした.

最初の試行結果は，図 2 に示すようにSASユーザー総会のトップページに掲載され，Excel化した論文のタイトル・著者名・年度・掲載ページから論文の検索ができるようになった．

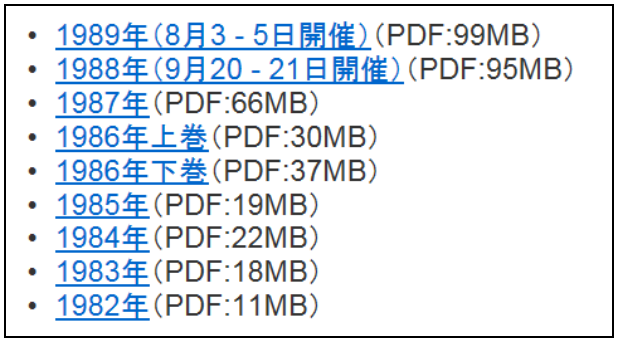


図 2　SAS ユーザー会のWEBのトップページのダウンロード画面

掲載後の外部の検索エンジンによって著者名順のExcelのリストが検索され，それをダウンロードして，Excel の検索機能を使って年とページを入手できるようになった．ただし，品質は不十分であり，サイズも 1988 年以後はページ数も多く 100Mバイト近くなり不満足な結果であった．品質上の問題は「かぶり」であった．WEB上の 1982 年論文集をダウンロードし 1－2 ページの「日本 SAS ユーザー会会則」拡大してみると，文字の周りに多くの点状のかぶりが見いだされる．このかぶりは，大きな文字の場合には印刷すると気になくなるが，SAS コード，結果の出力など小さな文字の場合には，判読不能となってしまう．

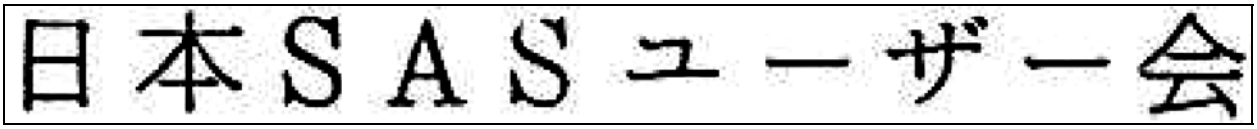


図 3　読み取り設定（グレイ，200 dpi）でのかぶり

1982 年の論文集は 84 ページで 11Mバイトなので，1 ページあたり 130 k バイトとなっている．ページ数がこの程度ならば「良し」としたかったのであるが，1989 年の論文集は 50 ページ，全体で 99Mバイトとなりダウンロードに躊躇するサイズとなってしまう．これは，スキャナー（Scan Snap S1500）での読み取り設定を，グレーで 200 dpi としていたためであった．

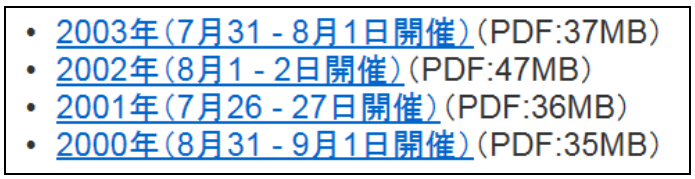


図 4　読み取り設定を（白黒，600 dpi）でサイズの縮小

そこで白黒の 600 dpi とすることにより品質の向上とサイズの削減が図られことになり，SAS ユーザー総会のトップページに図 4 に示すようなダウンロードページに示すようにサイズを 3 分の 1 にしつつ，品質の向上が図られた．

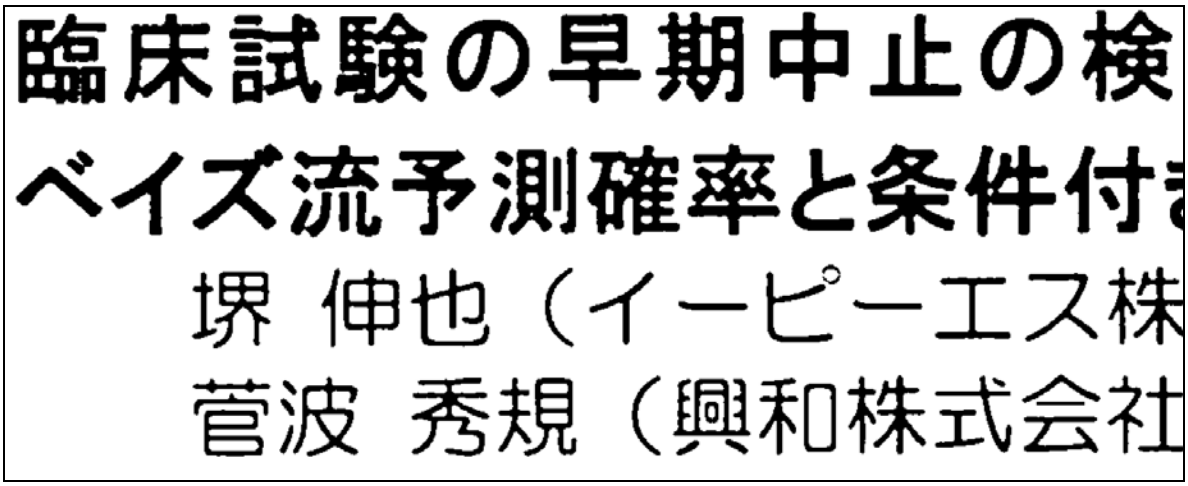


図 5　かぶりの消去（2003 年の目次のトップの拡大，白黒 600dpi）の確認

さらなる PDF のサイズの縮小を図るために Adobe Acrobat で最適化を実施すると 2003 年の論文集で 37 M バイトであったのを 9 M バイトと 4 分の 1 に縮小しつつ 図 6 に示すように拡大すると文字の輪郭にギザギザが生じているが，印刷した時にはほとんど判別がつかないことが確認された．また，印刷物では，文字が小さくて判読できにくいい場合でも，PFD を拡大して画面上で読み取れることも確認した．

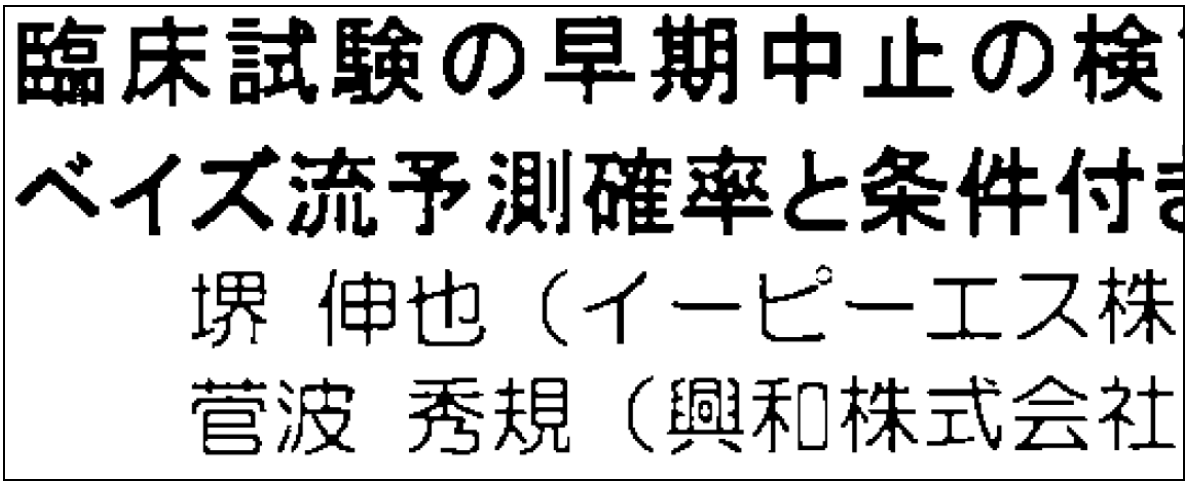


図 6　最適化後の品質

Adobe Acrobat によりイメージの文字を認識する OCR 機能を使い 32 年分の論文集の一括処理を行った．OCR による自動読み取りの結果を図 7 に示すが，そこそこの品質であり，2003 年の論文集で 5 M バイトの増え全体で 14 M バイトとなった．

臨床試験の早期中止の検
ベイズ流予測確率と条件付き
i 界仰也(イーピーエス株
菅波秀規(興和株式会社

図 7　OCR による文字認識の結果（堺信也が i 界仰也と文字化け）

3. 32 年分の論文集の目次の作成

年度ごとの論文集の目次について QCR 専用のソフト e.Typist を用いて文字化し，表 1 に示すように Excel に統合整理した．共著者がある場合には，所属が同じならばカンマ区切で入力し，所属が異なる場合には論文番号 11-12 で示すように新たな行とした．全体で 1,707 行，1,377 文献となった．

表 1 年次別タイトル一覧

論文番号	年度	年度番号	タイトル	名前	所属	開始	終	リンク
1	1982	1	日本SASユーザー会(SUGI-J)会則	日本SASユーザー会	日本SASユーザー会	1	2	SUGJ1982.pdf
2	1982	2	SAS導入の諸問題	高島邦彰	いすゞ自動車㈱	3	6	SUGJ1982.pdf
3	1982	3	SASの教育利用	雄山真弓	関西学院大学情報処理センター	7	10	SUGJ1982.pdf
4	1982	4	SAS/GR, APHと地図情報77	高橋均, 河津隆昭	国際航業㈱	11	14	SUGJ1982.pdf
5	1982	5	駿台電算専門学校でのSASの利用について	穂積和子, 須藤恵子	駿台電算専門学校	15	18	SUGJ1982.pdf
6	1982	6	SAS/GRAPHのXYプロッター，日本本語ラインプリンター等への出力について	福田正一	名古屋大学大型計算機センター	19	30	SUGJ1982.pdf
7	1982	7	SASと人口研究	小川真宏	日本大学人口研究所	31	38	SUGJ1982.pdf
8	1982	8	日立ソフトウェアエンジニアリング㈱におけるSASの導入	辻勝久	日立ソフトウェアエンジニアリング㈱	39	42	SUGJ1982.pdf
9	1982	9	SASの導入背景と利用について	北原精二	㈱富士銀行コンピューターサービスー	43	44	SUGJ1982.pdf
10	1982	10	SAS導入とその利用	長田博一	三菱化成工業㈱	45	48	SUGJ1982.pdf
11	1982	11	SASによる大麦データベースの試作	菅原秀明	理化学研究所	49	49	SUGJ1982.pdf
12		12		小西猛朗	岡山大学			SUGJ1982.pdf
：								
1703	2013	75	ミクロデータ分析，教育用擬似ミクロデータを用いた収入・消費傾向の考察	冨里遼太，土生敏明，米倉孝俊	大鵬薬品工業株式会社	545	548	SUGJ2013.pdf
1704	2013	76	ミクロデータ分析，シニア世代の消費特徴分析	中島貴之	株式会社データフォーシーズ	549	554	SUGJ2013.pdf
1705	2013	77	JMPClinicalにおけるCDISCデータの解析について	大津洋	東京大学大学院特任研究員	555	560	SUGJ2013.pdf
1706		78		山口拓洋	ー			SUGJ2013.pdf
1707	2013	79	索引2013	索引	索引	561	566	SUGJ2013.pdf

著者名別の文献リスト作成のために，カンマ区切りの名前を別々の変数として切り出し，行ごとの転置機能により 1 人ごとのファイルを作成し, 表 2 に示すように名前順の 2,454 人分のリストを作成した．さらに，名前順の文献リストと，名前の頻度順のリストを併合し，名前の頻度順の文献リストも作成し，検索の便宜を図ることにした．

表 2　名前順の文献リスト

論文番号	年度番号	名前	筆頭or共著	所属	タイトル	年	開始	終	リンク
170	23	A.C.B.Richardson	2	米国環境保護庁	米国職業被曝解析	1988	91	94	SUGJ1988.pdf
170	23	A.Wolbarst	2	米国環境保護庁	米国職業被曝解析	1988	91	94	SUGJ1988.pdf
1479	35	ARMAN BIDARB A	1	東京国際大学	Poverty Mapping: Case Study of Iran	2009	289	298	SUGJ2009.pdf
：									
1403	36	マヘシュ クマル ス	1	東京国際大学／ネパール中	ミクロ統計特別セッション、ネパールにおける貧困と不平	2007	330	333	SUGJ2007.pdf
993	36	ラーマチャンドラン	1	サティヤムコンピュータサービ	ウェブマイニングー競合優位性への道ー	2001	277	286	SUGJ2001.pdf
1401	34	ラクソノ アナン	1	東京国際大学／インドネシア	ミクロ統計特別セッション、貧困対策のための非貨幣的指	2007	304	317	SUGJ2007.pdf
1327	22	口羽文	1	東京大学/日本臨床腫瘍研究	nestedケース・コントロールデザインにおける擬似尤度に	2006	171	180	SUGJ2006.pdf
1208	25	阿部いくみ	1	三菱ウェルファーマ株式会社	前臨床実験データの統計解析をいかに検証するのか，前	2004	157	158	SUGJ2004.pdf
567	42	阿部まさ子	1	マリオン・メレル・ダウ株式会	Windows版SASのPCネットワークへの導入経験	1994	359	360	SUGJ1994.pdf
：									
1303	53	繆青	1	兵庫県立大学	JMPを活用した住民意識調査データに基づく行政課題の	2005	425	438	SUGJ2005.pdf
113	14	齊藤博	1	ヘキストジャパン㈱	HP3000によるホスト・システムの利用形態	1987	75	80	SUGJ1987.pdf
1008	51	翟国方	1	ダイナボット株式会社	データマイニング技法による生活習慣病のリスクファクタ	2001	407	416	SUGJ2001.pdf

1982 年から 2013 年の 32 年間について論文集の頻度を 図 8 に示す．1986 年から 2006 年にかけて，論文数は 40 件以上であったが，2007 年から 2012 年にかけて文献数の落ち込みがあり，SAS ユーザー総会の活動が縮小傾向となっていた．2013 年には 50 件と盛り返したことが読み取れる．

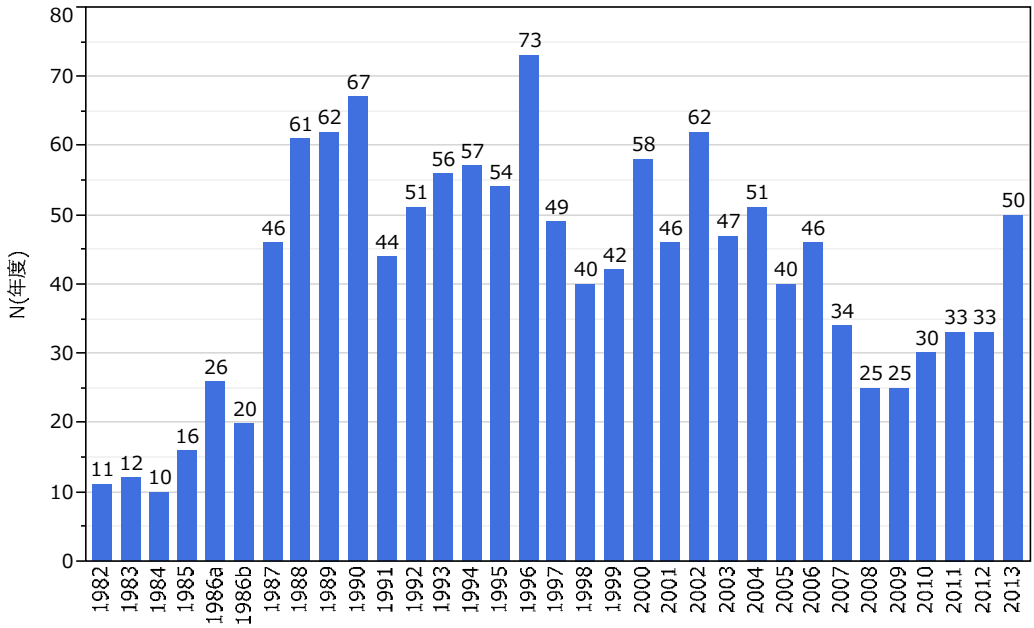


図 8　論文集の年度別頻度表

論文集には，1,332 人の発表者がおり，複数回の発表を含めると延べ 2,454 人であった．発表頻度の多い人もいれば，1 回のみの人たちもいる．共著も含めて 10 回以上，5 回以上，3 回以上，2 回，1 回の人数，論文総数，その割合などを 表 3 に示す．発表件数が 1 人で 54 回ものモンスター浜田知久馬氏もいれば，1 回しか発表していない人も多数いることがわかる．発表件数が 10 回以上は 21 人おり，論文件数では 365 報，14.9%であるが，1 回のみの人が 895 人で 36.5%と多くを占めていた．

表 3　発表者に関する統計

発表回数	人数	論文数	比率%
54回～10回	21	365	14.9
9回～5回	52	307	12.5
4回～3回	133	445	18.1
2回	221	442	18.0
1回	895	895	36.5
計	1322	2454	100.0

4. 文献検索の事例

表 1 に示した文献リストから所属が異なる共著者の行を削除して，SAS のプロシジャで“GLM”が含まれる文献リストを 表 4，“MIXED”が含まれる文献リストを 表 5 に示す．ある興味を持つ SAS のプロシジャについて調べたいときに，過去の SAS ユーザー総会の該当論文の検索が容易に行える．

浜田知久馬氏が筆頭著者である文献を 表 2 に示した発表者順のリストから検索した結果を 表 6 に示す．1992 年から 2013 年まで毎年欠かさずに筆頭著者として論文集に登場していることが見出される．

表 4 “GLM”が含まれる文献リスト

論文番号	年度	タイトル	名前	開始	終
44	1985	COX回帰プロシジャPHGLM使用経験上のある問題点について	森川敏彦	37	40
85	1986b	SASのGLMによる実験テークの解析	高橋行雄	35	52
131	1987	GLMの臨床試験データ解折への応用	高橋行雄, 藤丸清志	163	170
162	1988	GLMによる実験データ解析入門	高橋行雄	61	62
292	1990	Cox比例ハザードモデル(PHGLM)で求めた癌相対リスクと生荏期間の検証	大槻成章, 山縣清壮, 岡田頴一	15	22
362	1990	PROCGLMを用いた繰り返し測定データの解析	折笠秀樹	503	504
363	1991	Multiple Slopes Model(PROC GLM)による共分散分析の解釈	澤淳悟	1	8
733	1997	耐糖能障害・糖尿病改善に及ぼす要因の解析 -- GLMによる三元配置分散分析--	青野裕士, 小澤秀樹, 斉藤功, 池辺淑子	117	128
742	1997	各種の実験デザインにおける PROC GLM, PROC MIXED の利用	角元慶二	137	142
882	1999	PROC GLM及びPROC IMLを用いた3期3剤クロスオーバーデザイン(直交ラテン方格)の解析	石川靖	415	428
1215	2004	SAS/STAT GLMプロシジャの平方和計算の基礎	柴山忠雄	227	234
1349	2006	SAS/STATGLMプロシジャの演習－Excel 表示応答分解－	柴山忠雄	321	328
1485	2010	臨床試験データへの GLMSELECT procedureの適用	横溝孝明	63	74
1586	2012	GLM と MIXED による2剤2期クロスオーバーデザインの解析ー再考	斎藤和宏	137	150

表 5 “MIXED”が含まれる文献リスト

論文番号	年度	タイトル	名前	開始	終
469	1993	An Introductionto Mixed Model with the SAS/STAT MIXED Prccedure	Russell Wolfinger	35	42
546	1994	各種分割実験モデルに対するMIXEDプロシジャの活用	高橋行雄	183	202
669	1996	PROC MIXED入門	岸本淳司	179	198
742	1997	各種の実験デザインにおける PROC GLM, PROC MIXED の利用	角元慶二	137	142
751	1997	SAS MIXEDモデルを用いた成長曲線分析とその応用	李聖熙, 大竹正徳	183	188
874	1999	マルチレベル分析による生活満足度の分析 — SAS PROC MIXEDを用いて —	中田知生	349	360
900	2000	V8のODSによる総括報告書の電子化－関西プロジェクトー, その5.Model-based解析結果の要約(MIXEDプロシジャを例として)	伊藤要二	103	110
1045	2002	MIXEDプロシジャを用いた反復測定データの解析	菅波秀規	149	158
1098	2002	NLMIXED プロシジャを用いた項目反応理論モデルのパラメータ推定	伊藤陽一	485	496
1137	2003	MIXEDプロシジャを用いた線形混合効果モデルの交互作用の指定方法	寒水孝司	141	150
1165	2003	NLMIXEDプロシジャーを用いたItem Response Modelのシミュレーション	板東説也	361	368
1298	2005	MIXEDプロシジャを用いたメタ回帰	長谷川千尋, 渡部恵, 浜田知久馬	381	390
1330	2006	NLMIXEDプロシジャによるbreakpoint指数分布のあてはめ	浅野淳一, 浜田知久馬	191	202
1398	2007	メタアナリシスの功罪 －MIXED プロシジャによるメタアナリシスと公表バイアスへの対応	浜田知久馬	262	283
1586	2012	GLM と MIXED による2剤2期クロスオーバーデザインの解析ー再考	斎藤和宏	137	150
1629	2013	NLMIXED プロシジャを用いた生存時間解析	伊藤要二	73	82
1630	2013	NLMIXED プロシジャ紹介 PK 解析及び生存時間解析への応用	小林聡晃	83	96

表 6 浜田知久馬氏の SAS ユーザー総会における貢献

論文番号	年度	所属	タイトル	開始	終
460	1992	武田薬品工業㈱	MULTTESTプロシジャの紹介	357	370
503	1993	東京大学	SASによる生存時間解析	337	340
587	1994	東京大学	SASによる条件付きロジスティック回帰	527	540
613	1995	東京大学	SASによるメタアナリシス	241	254
687	1996	東京大学	SASによる用量相関性の解析	331	346
724	1997	東京大学	SASによる正確 (exact) な検定	17	34
837	1998	東京大学医学部	SASによる信頼区間の計算	375	394
838	1999	東京大学医学部	MULTTEST Q&A	3	18
876	1999	東京大学医学部	Separate-ranking型ノンパラ多重比較	383	390
891	2000	京都大学	V8におけるLOGISTIGの機能拡張	13	38
982	2001	京都大学	SAS V.8 における正確な推測とシミュレーションによる近似法	165	188
1043	2002	東京理科大学	V.8 における生存時間解析関連プロシジャの機能拡張	111	138
1129	2003	東京理科大学	生存時間解析における症例数設計	73	98
1211	2004	東京理科大学	SASV9のTPHREGを用いたメタアナリシス	165	194
1266	2005	東京理科大学	POWERプロシジャによる症例数設計	127	152
1317	2006	東京理科大学	ロジスティック回帰による推測(V.9LOG!STICプロシジャの機能拡張)	81	106
1398	2007	東京理科大学	メタアナリシスの功罪 －MIXED プロシジャによるメタアナリシスと公表バイアスへの対応	262	283
1434	2008	東京理科大学	SAS によるコクラン・アミテージ（Cochran-Armitage）検定	165	202
1480	2009	東京理科大学大学院	SASによる共分散分析	301	337
1492	2010	東京理科大学大学院	SASによる中間解析のデザインと解析	111	182
1524	2011	東京理科大学大学院教授）	生存時間解析入門「生存時間解析のミステリーをひも解く」	3	46
1576	2012	東京理科大学大学院教授	SASによる2値データの解析「ここまでできるFREQプロシジャV.9.3	3	58
1628	2013	東京理科大学大学院教授	SAS 生存時間解析プロシジャの最新の機能拡張	3	72

2014 年の SAS ユーザー総会では，企画セッションで「大学と企業における統計教育と SAS」が開催される．そこで，タイトルに“教育”が含まれる文献リストを作成してみた．第 1 回目，1982 年の雄山真弓氏による「SAS の教育利用」から始まり，その後も綿々と続いていることがわかる．

表 7 “教育”が含まれる文献リスト

論文番号	年度	タイトル	名前	開始	終
3	1982	SASの教育利用	雄山真弓	7	10
52	1985	第3WG（エンドユーザー教育研究グループ）	エンドユーザー教育研	71	71
74	1986a	情報処理教育におけるSASの利用 －文科系大学における一般教育課目での展開－	三浦協一	109	114
76	1986a	統計学教育とSAS	市川伸一	119	120
96	1986b	SASにおける教育システムの開発	佐藤栄里	101	106
104	1987	国際金融教育シミュレーションとSAS事例	川崎章弘	25	28
108	1987	ユーザー教育の手段としてのSAS/CBT	白石典義	43	48
110	1987	国際的機構におけるSASユーザー教育	グラトン ミサコ	59	64
177	1988	大学教育でのSASの実践例	武藤宣道，神田範昭，	125	132
178	1988	情報処理教育のあり方について	二宮正司	133	138
179	1988	SAS教育と認知カウンセリング	市川伸一	139	142
180	1988	社会科学教育とSAS	竹中治	143	146
182	1988	文科系学生に対するSAS利用教育	金井浩，静谷啓樹，川	151	158
327	1990	アイオワ州立大学統計学科におけるSAS教育およびSASの利用	布能英一郎	233	250
328	1990	社会科学専門教育と情報処理カリキュラム	竹中治	251	256
330	1990	教育心理学専攻学生に対するSAS教育	森際孝司	265	274
331	1990	PC-SASを利用した社会調査に関する大学教育	川上和久東	275	276
332	1990	理工系大学におけるSASによる統計教育の試み	山本英二	277	280
366	1991	バージニア州立大学医学部生物統計学科における統計学教育とSASの利用	大槻成章	17	24
382	1991	SAS/AFを利用した統計教育システム	東勲，能川賢一	107	110
522	1993	PC版SASによる情報処理教育	長野祐弘	453	454
523	1993	UNIX版SASシステムによる経済系情報処理教育(大阪大学経済学部での実施例)	田中克明	455	458
524	1993	歯科保健情繍生教育におけるUNIX版SASシステムの利用事列	松久保隆，大川由一，	459	464
525	1993	SASを使った統計教育 知的ツールとしてのパッケージ統計学	高橋伸夫	465	474
625	1995	利用者自身によるデータ活用のためのSAS教育の展開	八木章，高橋和子	317	326
878	1999	臨床疫学教育におけるSASの役割	縣俊彦，清水英佑，田	391	394
885	1999	企業における教育研修の評価と改善	陶山博太，伊藤洋子，	461	470
1009	2001	神戸商科大学におけるSASシステムを利用した統計・情報処理教育の現状と展望	川向肇，有馬昌宏，古	417	424
1031	2002	医薬特別セッション:JMPによる副作用データマイニング，JMP4Jによるロジスティック回帰モデルの教育－併用薬剤の種類，有害事象の種類別の探索的解析－	澤田克彦	81	90
1035	2002	医薬特別セッション:JMPによる副作用データマイニング，JMP4Jを使用した有害事象の生存時間解析の教育	西山智	91	98
1052	2002	実験計画法の学部内一般教育	柴山忠雄	185	192
1140	2003	CROにおけるSASプログラマの育成教育	竹田眞，佐藤智美	161	166
1171	2003	看護系大学における疫学・生物統計学教育の実態調査	田中司朗	391	400
1183	2003	SAS/GRAPH入門 ～ 社内における教育研修事例 ～	林行和，	477	488
1254	2005	臨床開発のためのSASプログラミング教育カリキュラムの開発と実践 ～統計解析業務を題材に～	山口孝一，林行和，平	13	22
1345	2006	SASを使った数値計算・統計処理教育プログラム	作花一志，南野公彦	297	308
1346	2006	テユートリアル教育(情報科学演習)における学習行動の類似性に関する定量分析	安田晃，平野章二，阿	309	320
1620	2012	統計教育と統計ソフトの共生	新村秀一	339	348
1642	2013	(財)日本科学技術連盟における「臨床試験セミナー統計手法専門コース」と SAS 教育	池田敏広	175	182
1682	2013	SASを用いた医薬品開発の統計解析担当者に対する CDISC の社内教育	浅見由美子，小山暢之	423	438
1703	2013	ミクロデータ分析，教育用擬似ミクロデータを用いた収入・消費傾向の考察	冨里遼太，土生敏明，	545	548

5. まとめ

SAS ユーザー総会の 32 年間にわたる活動の青果物である論文集をすべて電子化し，無料公開を目標に関係各位の協力のもとに実現した．SAS を活用し新たなチャレンジをしようとしている人たちが，これまでの熱狂的かつギルド集団的な活動の成果を踏まえ，さらなる SAS ユーザー総会の発展に積極的な関与をお願いしたい．また，無料公開することにより，多くの人たちが SAS のパワーを認識し，新たな SAS ユーザーとして活躍されんことを期待している．

金融／経済／システム

与信モデル構築

小野 潔　　松澤 一徳
株式会社インテック　金融ソリューションサービス事業本部

Credit Model of Developmennt for Data Mining

Kiyoshi Ono , Kazunori Matsuzawa
INTEC Inc. Business Solutions Developmennt Division

要旨

最近、貸金業法の総量規制の施行により、金融機関の与信モデルを見直す動きが始まっている。銀行等の金融機関が、従来、加盟できなかった個人信用情報センター(JICC)に加盟できるようになったためである。消費者金融会社の無担保ローン・モデルは、従来モデルよりも高精度であり、かつ安定的である。さらに金融庁が住宅ローンの収益性について言及しており、与信モデルの融資判定にも収益性の考慮が必要になりつつある。

金融機関の与信モデルの構築プロセスは一般モデルと同様にKDD (Knowledge Discovery in Databases)プロセスである。しかし与信モデルには"個人信用情報"、"精度を高める組合せモデル"、"モデル格付"、"判定マトリックス"、"AVR 領域"、"個別審査ルール"などの特徴がある。

本報告では、与信モデルの見直しの背景と、SAS/EnterpriseMinerを用いた与信モデルの構築法と特徴を報告する。また外部環境からの与信モデルのSASプログラムの起動法やSASマクロの代入法を解説する。

キーワード: 与信モデル　個人信用情報センター　データマイニング　マイニング・ツール
SAS/Enterprise Miner　ハイブリッド・モデル　アンサンブル・モデル
モデル格付　判定マトリックス　AVR 領域

1. はじめに

与信モデルは、住宅ローン、キャッシング、クレジット、マイカーローン等の融資を統計学に基づいて判定する。一方、審査システムは、申込の受付、保証会社や外部の個人信用情報機関とのデータ通信、決裁に至るまでの稟議/回覧、営業店からの照会などの審査工程を自動化し、審査担当者を補助するシステムである。自動審査システムは審査システムに与信モデルを組み込み、融資判定を自動化したものである。自動審査システムの導入メリットは、①判定の均一化(審査担当者による判定結果のばらつき防止)、②審査の業務時間の短縮、③リスクコントロールがある。

金融機関の与信モデルの特徴は、会社の保有するデータだけでなく①外部の情報ベンダーのデータも利用すること、②審査担当者が理解しやすい分析手法を採用すること、③倒産した場合の損失額は大きいので、より精度を高める組合せモデルを採用すること、④倒産率(≒スコア値)からモデル格付を確定し、判定マトリックスからAVR 領域[1]を決定する点があげられる。

[1] A 領域:モデルにより自動的に承認する領域、V 領域:審査担当者により判定する領域、R 領域:モデルにより自動的に謝絶する領域

与信モデルは、初期与信モデルと途上与信モデルがあるが、本報告では、SAS/EnterpriseMiner Ver12.0 を利用して初期与信与信モデルの構築を報告する。また自動審査システムは通常、Java 等のプログラミング言語で開発されため、本報告では実務上の観点にたち、外部環境からSASシステムを起動し、かつマクロ変数による引数渡しを説明する。

2. 背景

与信モデルの構築は、都市銀行で2000年頃から、地方銀行で2006年頃から導入が始まった。現在、与信モデルはメガバンク、大部分の地方銀行に多数導入されている。最近、貸金業法の総量規制の施行により、金融機関の与信モデルを見直す動きが始まった。理由の一つは、貸金業法の総量規制実施の影響で、銀行等の金融機関が従来、加盟できなかった個人信用情報センター(株式会社日本信用情報機構:JICC)に加盟できることになったためである。JICC は全国の消費者金融会社が加盟しており、加盟会社はリアルの無担保ローンに関する個人信用情報を入手できる。消費者金融会社の無担保ローン・モデルは、従来モデルよりも高精度であり、かつ安定的である。銀行等の金融機関も 2012 年から情報を入手ができるようになり、現在、個人信用情報を利用した無担保ローン・モデルが期待される。

もう一つの理由は、近年、金融庁や日本銀行が住宅ローンの与信判断に収益性を求めている点である。従来の住宅ローンの融資判定でも収益を加味されていたが、残高、金利、事務経費から算出する単純な収益であった。しかし近年の住宅ローンの競争激化に伴う金利低下により、金融庁は金融機関に対して、中途解約や経年解約や金利変更に伴う収益を考慮するように指導を始めた。そのため、住宅ローン・モデルに収益性を含めて融資判定させる動きが発生している。ただ金融庁・日本銀行が求めている住宅ローンの収益は、個人の生涯収益であり、理論が未だに完全に確立されていない。

3. (参考) 個人信用情報センター

日本の個人信用情報センターは、①銀行および子会社が加盟する全国銀行個人信用情報センター(略称:KSC)、②信販会社(≒クレジット会社)が加盟する株式会社シー・アシ・シー(略称:CIC)、③消費者金融会社が加盟する株式会社日本信用情報機構(略称:JICC)がある。従来は各情報センターは、異業種の加盟を認めなかった。

貸金業法は2006年12月に成立、システム対応の準備期間が必要のため、2010年6月に総量規制を含むすべての規定が施行された。その結果、新たな貸付けの申込みを受けた場合、貸金業者は指定信用情報機関が保有する個人信用情報を使用し、他の貸金業者からの借入残高を調査することが法令で定められた。

総量規制の実施に伴い、個人の借入総額が必要になり、指定信用情報機関制度が導入された。JICC と CIC は 2012 年に貸金業法に基づく「指定信用機関」として内閣総理大臣より指定を受けた。それに伴い銀行などの金融機関も、消費者金融などの貸金業者と同じようにJICC への加盟が認めれ、借入額や件数などを開示請求できるようになった。今では82銀行(銀行の65%、表1参照)がJICC に加盟し、ノンバンク、消費者金融などの個人の借入情報を見ることが可能になった。

貸金業法の総量規制は個人の借入総額が、原則、年収等の 1/3 までに制限する。対象は「個人向け貸付け」であり、個人の金の借入れである。ただ個人が事業用資金として借入れる場合は、総量規制の対象にならず、クレジットカードを使った商品購入は貸金業法の対象外である。また総量規制は貸金業者か

らの貸入れを対象としており、銀行からの借入を対象外である。

	2014/5/31	2009/5/31
都市銀行	5	6
信託銀行	3	4
地方銀行	64	64
第二地方銀行	41	44
その他銀行	13	12
合計	126	130

・「その他銀行」は、新生銀・ジャパンネット銀・セブン銀・楽天銀・新銀行東京・あおぞら銀・シティバンク銀・住信SBIネット銀・イオン銀・大和ネクスト銀・SBJ銀
・信託銀行は銀行子会社等を除く専業信託銀行に限定
・出典：ニッキン

表 1　日本の銀行数

信用情報センター（JICC）の個人信用情報の登録内容を表 2 に示す。

	内　容	登録期間
本人特定情報	氏名、生年月日、性別、住所、電話番号、勤務先、勤務先電話番号、運転免許証等の記号番号等	契約内容に関する情報等が登録されている期間
契約内容	登録会員名、契約の種類、契約日、貸付日、契約金額、貸付金額、保証額等	契約継続中及び完済日から5年を超えない期間
返済状況	入金日、入金予定日、残高金額、完済日、延滞等	契約継続中及び完済日から5年を超えない期間 (ただし、延滞情報については延滞継続中、延滞解消の事実に係る情報については当該事実の発生日から1年を超えない期間)
取引情報	債権回収、債務整理、保証履行、強制解約、破産申立、債権譲渡等	当該事実の発生日から5年を超えない期間 (ただし、債権譲渡の事実に係る情報については当該事実の発生日から1年を超えない期間)
申込み情報	本人を特定する情報（氏名、生年月日、電話番号及び運転免許証等の記号番号等）、並びに申込日及び申込商品種別等	申込日から6ヵ月を超えない期間
・・・・・	・・・・・	・・・・

表 2　信用情報の登録内容

4. 無担保ローンの与信モデルの特徴

無担保ローンの与信モデルでは、「個人信用情報を利用したモデル」と「利用しないモデル」では、予測精度とモデル安定性に大きな差が存在する。与信モデルの決定木分析（後述）では、倒産に相関が強い順にデータ項目のツリーが作成される。図 1 は両者のモデルを比較した概略ツリー図である。「左図が個人信用情報を利用したモデルのツリー図」、「右図が個人信用情報を利用しないモデルのツリー図」である。個人信用情報を利用した場合、特に **1, 2** 層に個人信用情報のデータ項目が現れる。つまり個人信用データを利用すれば、与信モデルの精度を向上が期待できる。

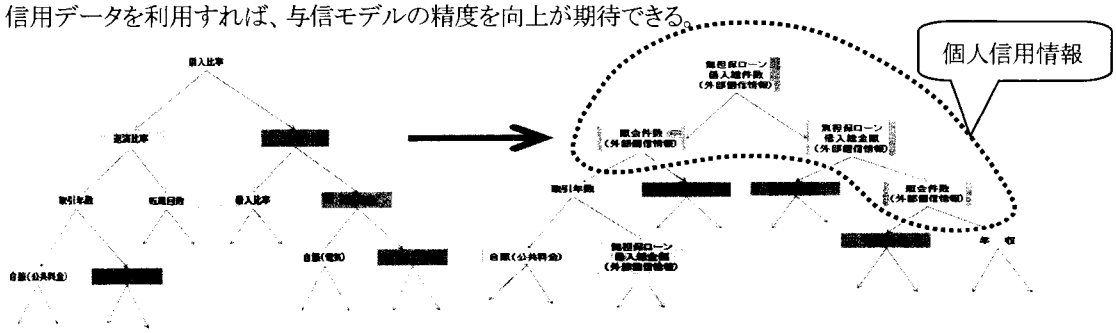


図 1　個人信用情報を含むモデル(左)と含まないモデル(右)のツリー図比較

なお銀行は個人信用情報センターのKSCから個人信用情報を入手できたが、無担保ローンに関してはマーケティング範囲も狭く、リアル性も少ないため、与信モデルの否認基準等に利用されてきた。

5. マイニングツール

与信モデルの構築はマイニングツール(SAS/EnterpriseMiner、IBM/Modular 等)を利用する。マイニングツールはKDDプロセス(Knowledge Discovery in Databases、データベースからの知識発見)を実装したものであり、プロセスを繰り返すことで、精度の高い知識が得られる。KDD プロセスは、①選択(selection)、②前処理 (preprocessing)、③変形 (transformation)、④データマイニング (data mining)、⑤解釈・評価 (interpretation / evaluation)、⑥ルール生成(rule)から成り立つ。KDDプロセスは分析だけはなく、マイニングの知識発見の一連作業を指す。マイニング・ツールは、KDD プロセスを実装し、モデルのスクラップ&ビルドをプログラムレスで実現した。現状ではツールなしでのモデル構築は考えられない。

6. 分析手法

本項では与信モデルによく使われる決定木分析を解説する。決定木分析は、データから専門家の知識(IF-THENルール)を抽出するための人工知能学の技法である。信頼区間や安定性に関する数値は算出しないが、ため、分析結果のIF-THENルールが人間にとってわかりやすい。

決定木分析は、データ属性を使ってグループ分類でき、分析結果は樹形(ツリー)図で表現される。決定木分析では、ツリーの分岐の属性順序や閾値を自動的に算出するので、分析者の恣意は排除される。ツリーの構造から「もし…ならば～である」という IF-THEN ルールを導出できる。決定木分析のメリットは IF-THEN ルールに直すことで、審査担当者の理解を得られやすくなる。デメリットは精度を向上させるには、ツリーの階層を増やすことになるが、5～7 階層に達すると、IF-THEN ルールが細かくなりすぎて、専門家でも全体像をつかめなくなる点である。

決定木分析の優秀な点は、属性の分割基準値に基づいて、分割属性の優先順位が決まることにある。分割基準値は、ルールが目的属性値の分布与える影響度合いを数値化したものである。基準値が小さいほど影響力が大きく、ツリーの最初の分割属性になる。

決定木分析分割基準には、「情報エントロピー値」、「GINI 基準値」、「カイ 2 乗値」という代表的な分割基準がある。データ集合 S に、j 個のカテゴリー値をもつ目標属性が存在し、集合 S 内にi個番目の値をもつデータがそれぞれ Xi(S)個(i=1, . . , j)あると仮定する。ルール R で S1 と S2 に 2 分割し、部分集合 S1 内のi番目の値の分布比率を Pi(Si)=Xi(S1)/|S1|とすると、各分割基準を表 3 のように定義できる。

決定木の種類	分割基準値	定 義 式
C5.0, C4.5	情報エントロピー値	$Ent(R) = Ent(x(S_1))$ $= -\sum_{i=1}^{k} p_i(S)\log\ p_i(S) - (-\frac{\|S_1\|}{\|S\|}\sum_{i=1}^{k} p_i(S_1)\log\ p_i(S_1) - \frac{\|S_2\|}{\|S\|}\sum_{i=1}^{k} p_i(S_2)\log\ p_i(S_2))$
CART	GINI値	$Gini(R) = Gini(x(S_1))$ $= (1 - \sum_{i=1}^{k} p_i(S)^2) - \frac{\|S_1\|}{\|S\|}(1 - \sum_{i=1}^{k} p_i(S_1)^2) - \frac{\|S_2\|}{\|S\|}(1 - \sum_{i=1}^{k} p_i(S_2)^2)$
CHAID	カイ2乗値	$Chi(R) = Chi(x(S_1))$ $= \sum_{i=1}^{k} \frac{\|S_1\|(p_i(S_1) - p_i(S_1))^2 + \|S_2\|(p_i(S_2) - p_i(S_2))^2}{p_i(S)}$

表 3　分割基準の定義式

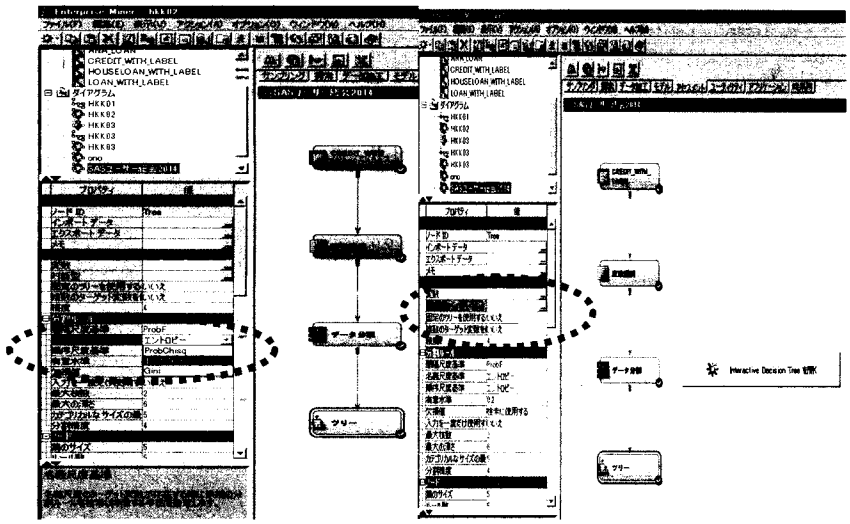


図 2　決定木の選択法(左)と対話モードへの移行(右)

SAS/EnterpriseMiner の分割基準の変更を図 2 左に示す。与信モデルの決定木分析では、一度ツリーの生成に成功したら、次にマニュアル操作に移行し、審査担当者のヒアリング結果に合わせて、ツリーを作りなおす。図 2 右図に対話モードへの移行法を示す。対話モードではデータ項目の優先順位変更や数値の分岐を恣意的にできるつくるが、一度、完成したモデルの精度にあまり影響しない。

7. 分析結果

外部の個人信用情報を“利用したモデル”と“利用しないモデル”を ROC 図(図 4)と累積リフト(図 5)に表示する。4 図と 5 図から、“個人信用情報を利用したモデル”は、優れた精度を有していることがわかる。

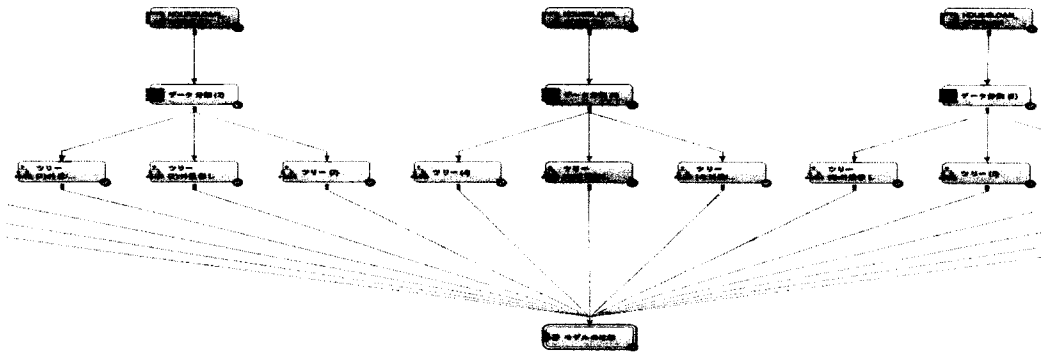


図 3　**SAS/EnterpriseMiner** のワークスペース上のモデル作成

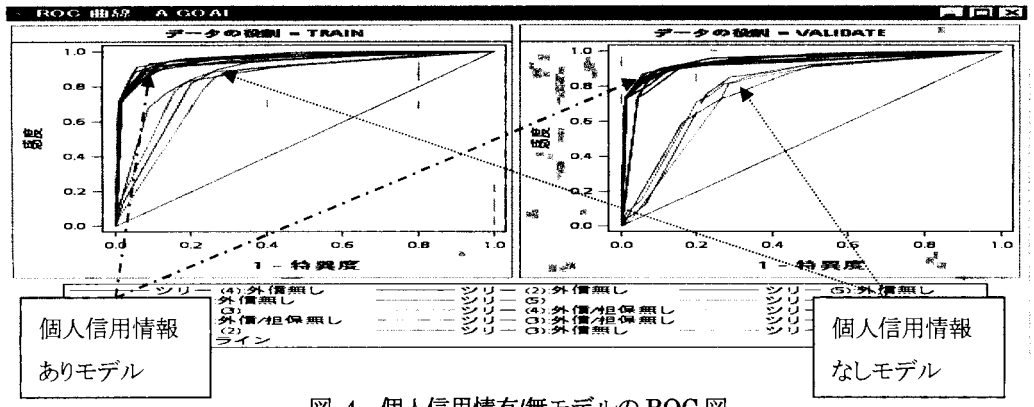


図 4　個人信用情有/無モデルの ROC 図

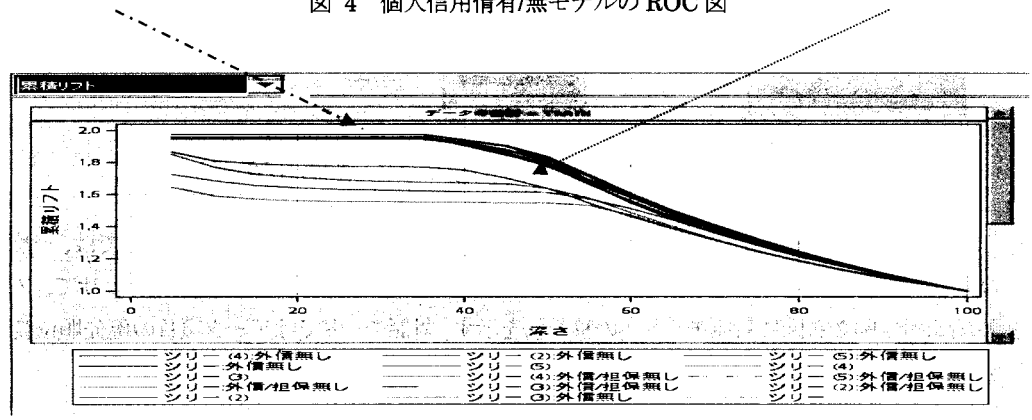


図 5　個人信用情有/無モデルの累積リフト図

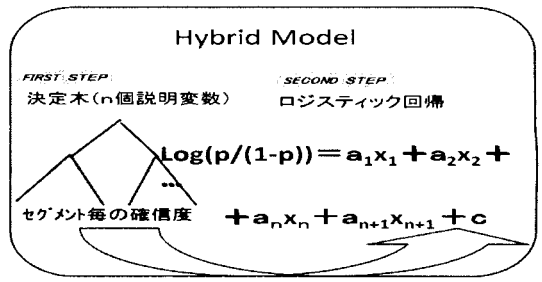
8.　与信モデルの分析手法とモデル

日本の金融機関の与信モデルでは、わかりやすい分析手法として、決定木、ロジスック回帰が採用されるケースが圧倒的に多い(MBR を採用されたケースも有り)。ニューラルネットはパラメーターチューニングに成功すれば、他の手法と比べても精度が高いため、米国の金融機関ではニューラルネットワークが採用されるケースも見受けられる。しかしニューラルネットワークは判定理由を明確に示すことができないため、審査担当者から理解を得ることがむずかしい。

与信モデルの構築では、倒産した場合の損失が大きいため、オーバーフィッチングしない限界までモデルの精度をあげる必要がある。そのため単独分析手法ではなく、複数の分析手法を組み合わせて使用する(本稿では分析手法の組合せ方をモデルと称する)。日本の与信モデルで採用され、実績のあるモデルは、“ハイブリッドモデル”、“アンサンブルモデル”、“カテゴリーフラグモデル”である。

8.1.　ハイブリッドモデル(2段階-直列型モデル)

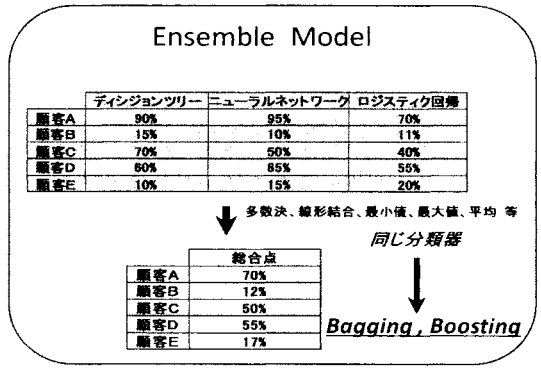
このモデルは 2 つの手法を直列に組み合わせる方法である。第1段階の分析手法で得られた倒産率を説明変数に追加する点がミソである。第2段階の分析では第1段階で得られた倒産率の寄



与度(説明力)が高いため、第2段階の分析手法は精度の微調整を行い、更に精度を高めていく。あたかも衛星ロケットが第1段ロケットで大気圏まで上昇し、第2段ロケットで衛星軌道の乗るために微調整の起動修正を行うようなものである。

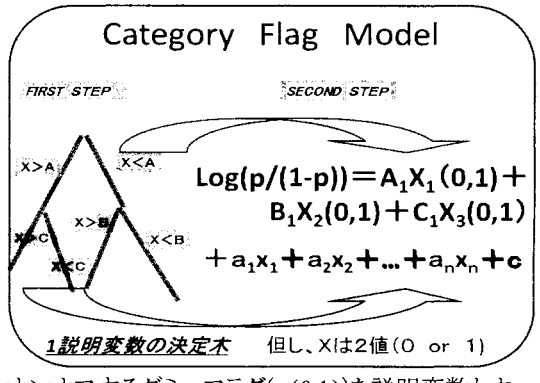
なお分析手法は多数存在するので、組合せ方は無数になるが、与信モデルでは最初に決定木分析を行い、次にロジスティック回帰分析を適用する。理由は決定木分析から得られるIF-THENルールが、現場の審査担当者にわかりやすいためである。しかも第2段階にロジスティック回帰分析を用いるため、最終推定モデルが一つの式で表すことでき、運用上の取り扱いが容易になる。

8.2. アンサンブルモデル(多数決-並列型モデル)



アンサンブルモデルは"あまり精度が高めない分析(弱い分類器)"を複数回行い、多数決で判定する方法である。複数の分析は「異なった分析手法」でも「ランダムサンプリングにしたデータに基づく同じ分析手法」を適用してもよい。つまりモデルを5個～100個程度作成し、各分析結果の倒産率の平均値、中央値等を倒産率の代表値とする。精度が高める分析、例えば前述のハイブリッドモデルを複数個作成しても、倒産率の差があまりないため、精度の改善が望めない。例えるならば、経験が浅い10人の審査担当者が同じ案件を審査し、多数決で判定するようなものである。もし全員がベテラン(ハイブリッドモデル)ならば、同一判定となり多数決にならない。

8.3. カテゴリーフラグモデル(離散型モデル)



金融機関のデータにはカテゴリデータ(例:職業、住居形態)と数値データ(例:給料、勤務年数)から成り立つ。このモデルは、すべてのデータをカテゴリーデータ(離散化)にすることで、分類器の判別力を高める方法である。ただ数値データの離散化は情報量の劣化を招きかねない。そのため、情報量の劣化を最小限にする閾値を選択する方法として、分割基準値を利用する。閾値が決まれば、カテゴリー範囲をオンオフするダミーフラグ(={0,1})を説明変数とするカテゴリー変数を追加する。この作業をすべての連続数値に繰り返し、最後にロジスティック回帰分析を適用する。

9. 個別ルールの制定

与信モデルは与信に関わるすべて事象をモデル化できない。与信モデルを簡単に言えば、データ項目の優先順位に重みをつけて、何らかの演算を行って倒産率を算出することである。つまり与信モデルの欠点は大数の法則にしたがうため、発生ケースが少ない特異な案件に対しては、有効性がない点にある。

この点を補完するために審査の個別ルールを作成する。個別ルールの発見方法は、①審査担当者

へのインタビューから経験則や知見を掘り起こす方法と、②自動的に作成した多数ルールの中から有効ルールの選別する方法がある。

本報告では SAS/Enterprise Miner を利用した②自動的にルールを作成する方法を説明する。この分析にはアソシエーション分析を利用する。アソシエーション分析は多量データからアイテム(ここではデータ項目)の関連性について自動的に抽出する方法である。

アソシエーション分析を利用すると、倒産事象に関連するイベント(データ項目)の組合せを抽出できる。具体的には目標変数の倒産に該当するデータのみを SAS データセットにあつめ、カテゴリー毎に 1 または 0 のフラグを作成し、SAS/Enterprise Miner のアソシエーションノードを適用する。ただアソシエーション分析から抽出される大量のアソシエーションルールのほとんどは意味がなく、その中から意味のある個別ルールを探す。SAS/Enterprise Miner では大量データを処理するため、IBM が開発したアプリオリというアルゴリズムを採用して短時間で処理する。

個別ルールを否認条件に含めるか、それとも独立した個別ルール・システムに含めるかは、得られたる内容による。個別ルールの取扱いは面倒なケースが多いため、可能ならばメインの分析、例えばツリーの枝葉に吸収できないかを考える。

10. モデル格付

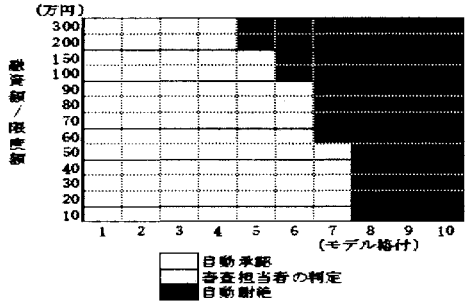
便宜上、本稿では与信モデルから算出する値を倒産率と述べていたが、与信モデルから算出される値はスコア値という。融資商品のデフォルト率は数%程度のため、倒産の件数が少ない。このデータ比率でモデル開発すると、倒産の特徴を読み取れなくなる恐れがある。そのため、モデル開発では、倒産件数を最大値とした、Black:White=1:1 のサンプリングを行う。その結果、与信モデルから算出されるスコア値の平均値は 50%になる。このスコア値を倒産率へ変換するには、①実際のデータから、あるいは②ベイズ統計学を利用する方法がある。

モデル格付の設定は、①スコア値で均等に 10 分割する方法と、②スコア分布の構成比の 10 分位をとる方法と、③スコア値を倒産率に変換し倒産率の基づき 10 分位を設定する方法がある。

11. 判定マトリックス

与信モデルは倒産率や格付を算出するが、融資判定には別ロジックが必要である。融資判定には、倒産率だけでなく、倒産した場合の損失額(≒回収額)あるいは外部個人信用情報を考慮する。そこで融資判定は、モデル格付けと回収金額((LTV:LoanToValue)が望ましい)あるいは専業件数(もしくは総額)等を軸とした判定マトリックス(次頁、図6参照)を利用する。判定マトリックスは AVR 領域に分割されている。AVR 領域は倒産率から導いた格付と回収金額による与信判断の意思決定マトリックスで表現する。要は AVR 領域の判定は、審査担当者が審査するグレー領域を特定することである。

最近、住宅ローンでは日銀・金融庁の指導により、収益を加味する要望が高い。そのため、判定マトリックスに収益性も含めた 3 次元として捉えるようになってきた。ただ金融庁・日本銀行の住宅ローンの収益は、単純に金利と残高と事務経費から求めるだけでなく、中途解約、経年解約、金利変動リスクも考慮し、他の金融商品の収益をふくめた生涯収益を理想としている。しかし、住宅ローンの収益モデルは理論的に未完成の部分があり、しかも長期わたる住宅ローンのすべて実務データが揃っている銀行はいない。実務では、収益計算の一部簡略化して計算せざるを得ないのが実情である。

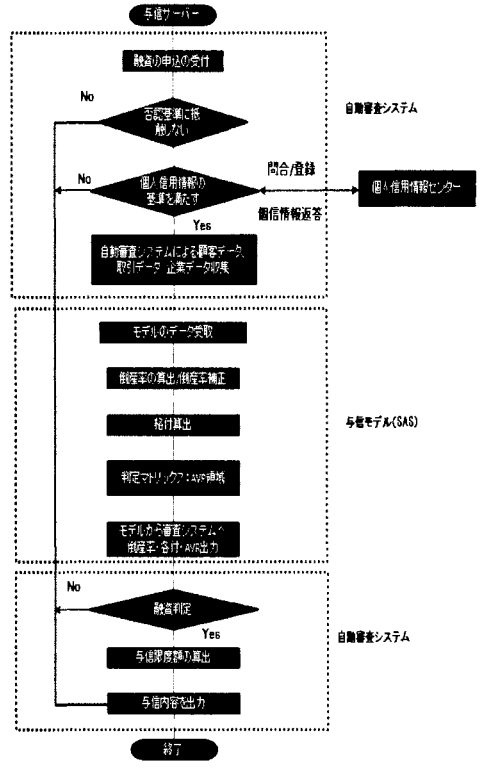


AVR領域	説　明
A(ACCEPT)領域	与信モデルにより自動的に判定を承認する領域 いわゆるWhite領域
V(REVIEW)領域	審査担当者により審査の承認・謝絶を判定する領域 いわゆるGray領域
R(REJECT)領域	与信モデルにより自動的に謝絶を判定する領域 いわゆるBlack領域

図 6　AVR 領域

12. 自動審査システムのフロー

自動審査システムは与信サーバーに構築される。自動審査システムのフローを左図に示す。まず案件は否認基準(行内のコンプライアン情報や延滞情報)でチェックする。次に個人信用情報センターへお客さまの個人信用情報の登録と問合せを行う。借入総件数や借入総額が否認基準に抵触していないかどうかを照合する。否認基準をパスした融資案件は次に行内の取引情報、勤め先の情報、その他情報を集められ、SAS の与信モデルが起動する。与信モデルではスコア値(倒産率)を算出し、モデル格付を確定させる。判定マトリックスで AVR 領域を判定し、自動審査システムへ出力する。自動審査システムでは AVR 領域に合わせて融資判定を行う。A/R 領域ならば自動判定で、V 領域ならば審査担当者が融資の判定を行う。



13. 外部プログラムから SAS プログラムの実行方法

審査システムは Java や C などのプログラミング言語で開発されるため、SAS プログラムの与信モデルは外部プログラムから起動する。本項では外部プログラムから SAS プログラムを実行する方法と、外部プログラムからの SAS マクロ変数の代入法を示す。

① Windows　Bat ファイルからの実効

コマンドライン(bat プログラム)を実行することで、開発 SAS コードを Window から起動できる。

【syntax】　c:¥(..SAS インストールパス..)¥sas.exe

-sysin c:¥(..実行する SAS プログラムのパス..)¥mypgm.sas

-config c:¥(..インストールした SAS の config ファイルのパス..)¥sasv9.cfg

【example】

パラメータ等	引数の意味
c:\program files\sas\sas 9.1\sas.exe	SASシステム実行
-sysin C:\mysas\programs\mypgm.sas	-sysin以下に実行するsasプログラムの場所を記載
-sysparm Tokyo	-sysparm Tokyo "と書くことでマクロ変数 &sysparm に 文字データ"Tokyo"を代入
-config c:\program files\sas\sas 9.1\sasv9.cfg	-config以下にsasのconfigファイルを記載

② powershell から bat ファイルを呼び出す方法

実務では、windows bat のみのバッチ開発は難しいため、powershell 等から前述のコマンドを記載した bat ファイルを呼び出す方法を採用する。

【syntax】 cmd /C "call c:¥(..SAS 実行用の bat ファイルパスを指定..)"

・・・・・・・・・・・・・・・・

14. おわりに

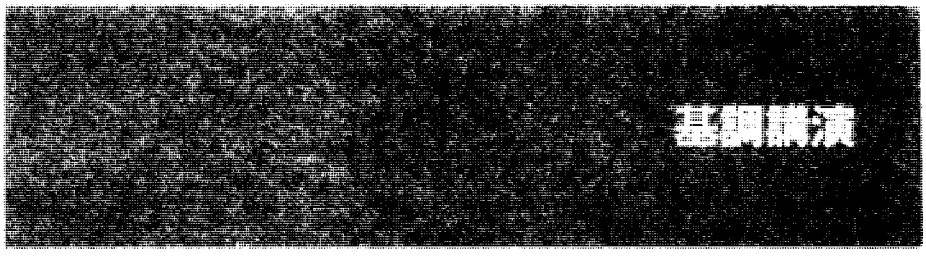
モデル構築には SAS/EntepriseMiner 等の優秀なマイニングツールが不可欠である。ただ与信モデルの構築は、単に精度のよいモデルを開発するだけでなく、現状の金融事情を把握し、例えば金融庁・日本銀行の方針に合わせたモデルを構築する必要がある。また審査担当者の理解できるようなモデルを作ることも重要である。本報告では一般的な与信モデルの構築法を述べたが、実際の開発にはデータクリーニング、恣意的データの混入、倒産データの不足、未完成な理論などの様々な壁が存在する。

以前は、所属する金融業界(銀行、信販会社、消費者金融会社等)により、与信モデルは相違していたが、今回のように同じ個人信用情報が使えるようになると、優れた与信モデルへ収束していくことになろう。現状はその過渡期にある。ただ銀行が消費者金融会社型のモデルを運用しても、不良債権の回収率が違うため、本報告で説明した収益を含めた判定マトリックが違うものとなり、運用面は同じにならない。与信モデルは、今後も所属する金融業界に合わせて、構築する必要があろう。

15. 参考文献

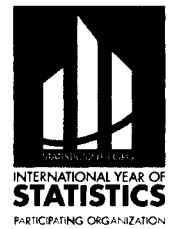
・Fayyad, U.M., Piaetsky-Shapiro, G. and Smyth, P.: From Data Mining to Knowlede Discovery: An Overview, Advances in Knowlege Discovery and Data Mining, pp.1-34, AAAI/MIT Press, 1996.
・Joseph P.Bigus, "Data Mining with Neural Networks", The McGraw-Hill Companies, 1996.
・コアブ・フロインド,ロバート・シャピリ,訳:阿倍直樹,"ブースティング入門",人工知能学会, vol.14 No.5,pp771-780,1999.
・小野潔,"データマイニングを利用した融資モデルの現状と課題",人工知能学会研究会資料 SIG-J-A004,pp49-54,2001.
・小野潔,"ハイブリッド・コンポーネントの構築", 第 20 回日本 SAS ユーザ会研究発表論文集,pp269-327,SAS Institute Japan,2001.
・河野浩之,"データベースからの知識発見の現状と動向"、人工知能学会, vol.12 No.4 ,pp.497-504,1997.
・寺野隆雄、"KDD ツールの動向と課題"、人工知能学会, vol.12 No.4 ,pp.521-521,1997.
・丸岡章,滝本英二,"オンライン予測",人工知能学会, vol.14 No.5,pp763-770,1999.
・小野潔,"マイニング・ツール選択のポイント",日経情報ストラテジー, vol.7, pp.56-59, 日経BP社,2000.
・小野潔,"データマイニングを利用した融資モデルの現状と課題",人工知能学会研究会資料 SIG-J-A004, p p 49-54,2001.
・J.R.キンラン、"AIによるデータ解析"、トッパン、1995.

・丹後俊郎、山岡和枝、高木晴良、“ロジスティク回帰分析”、朝倉書店、1996.
・大橋靖雄、浜田知久馬、“生存時間解析”、東京大学出版、1995.
・鈴木義一郎、“情報量基準による統計解析入門”、講談社、1995.
・エリザベス・メイズ、“クレジットスコアリング”、シグマベイスキャピタル、2001.
・大久保豊、尾藤剛、“ゼロからはじめる信用リスク管理”、きんざい、2011.
・本田義一郎、三森仁、“住宅ローンのマネジメントを高める”、きんざい、2004.
・日本銀行金融機構局、“住宅ローンのリスク管理”、2007.
http://www.boj.or.jp/research/brp/ron_2007/data/ron0703c.pdf
・日本銀行金融機構局、“住宅ローンのリスク・収益管理の一層の強化に向けて”、2011.
https://www.boj.or.jp/research/brp/ron_2011/ron111124b.htm/
・小島俊郎、“金融庁検査結果事例にみる住宅ローンビジネスの現状”、野村資本市場クォータリー2012Spring.
http://www.nicmr.com/nicmr/report/repo/2012/2012spr10.pdf


基礎講演

統計分析における「第三の変数」の功罪

成蹊大学理工学部情報科学科
教授　岩崎 学
iwasaki@st.seikei.ac.jp



自己紹介

1952年12月14日　静岡県浜松市生まれ

- 学会など
  - SASユーザー会　名誉会員
  - 統計関連学会連合：副理事長
  - 日本統計学会：代議員，前理事長（2期）
  - 日本計量生物学会：評議員
  - 日本行動計量学会：理事，編集委員
  - 応用統計学会：評議員
  - 合計：理事28期，評議員（代議員）34期
- 政府機関など
  - 消費者庁消費者委員会：専門委員
  - 医薬品医療機器総合機構：専門委員
  - 文部科学省，総務省，厚生労働省などの各種委員

要旨

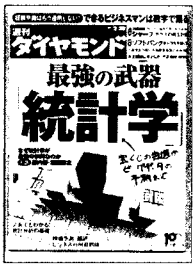
- 統計的データ解析では因果関係の確立が大きなテーマ
- ビッグデータ解析でも，将来に対する方略・戦略の立案では，何をすればどうなるかの正しい知識が必要
- 因果関係では，文字通り「原因変数」と「結果変数」があるが，それに加え「第三の変数」が重要な役割を果たすことが多い
- これらは，無視したり使い方を誤ったりすると結果に偏りをもたらす
- 本講演では，それら「第三の変数」の正しい使い方について，分かりやすく解説

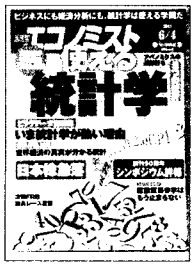
The Sexy Job

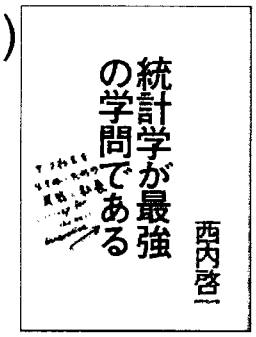
- Hal Varian on How the Web Challenges Managers (2009)
  - Google's chief economist
- I keep saying the sexy job in the next ten years will be statisticians.
- The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids.

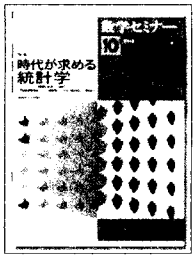
5

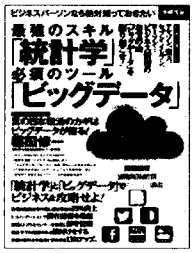
さまざまなマスコミで (2013)







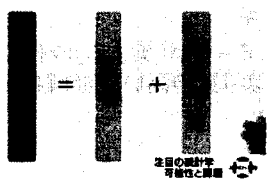




6

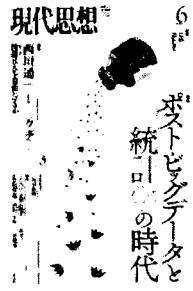
NHK でも

- 2013年7月3日（水）クローズアップ現代
- 数字のカラクリ・データの真実～統計学ブームのヒミツ～
- 視聴率：10.7%（関東地区）



7

現代思想 (2014年6月号)



特集:ポスト・ビッグデータと統計学の時代
【イントロダクション】
ビッグデータと統計学 ／ 竹内 啓
【討議】
情報(データ)は人を自由にするか ／ 西垣 通+ドミニク・チェン
【インパクト】
統計学にとって情報とは何か ／ 竹村彰通
ビッグデータブームを考える ／ 水田正弘
ビッグデータは科学を変えたか？ ／ 出口康夫
【インタビュー】
統計学は科学の文法である 水俣から福島まで、なぜ公害は繰り返されるのか ／ 津田敏秀
【統計学の現在】
統計的因果推論の考え方 ／ 岩崎 学
統計学・確率論の有効性とその限界 ／ 小島寛之
統計・実証主義・社会学的想像力 ／ 太郎丸 博
【データという問題】
ビッグデータの社会哲学的位相 ／ 大黒岳彦
「非有機的身体」の捕獲 膨脹する所与(データ)と新たな利潤(レント)源泉 ／ 長原 豊
工学的心身問題 ／ 西川アサキ+森脇紀彦
【ポスト・ビッグデータ社会のために】
生かさない〈生—政治〉の誕生 ビッグデータと「生存資源」の分配問題 ／ 柴田邦臣
「ネオ精神医学」を生み出した「トロイの木馬」:DSM アメリカにおける父殺しと科学への倒錯 ／ 樫村愛子
ビッグデータとビッグソサエティ ／ 和田伸一郎

8

日経産業新聞 (2014. 6.10)

- 統計解析最前線

- ビジネスの場で生かす統計解析
  - 欧米で当たり前の統計解析がなぜ日本企業で遅れているのか
  - 統計解析を企業利益につなげる人材の登用・育成がカギ
  - データが「集まる」時代こそより質の高い統計解析を

9

統計的データ解析の流れ

10年以上前のスライドだが

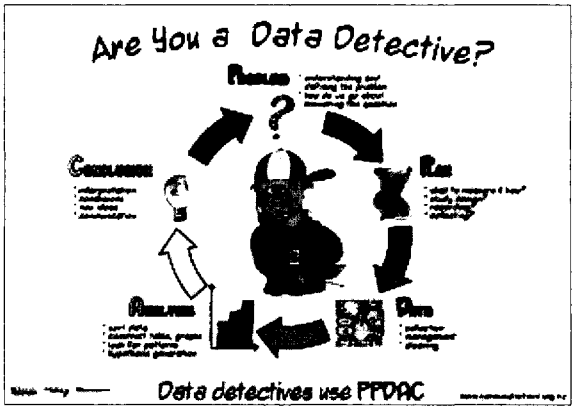
- 研究目的の設定
- データ収集法の立案：実験，観察研究，調査
- データの収集（モニタリング）
- データの電子化
- データのチェック（クリーニング），マージ
- データの集計とグラフ化（予備的検討）：記述統計
- 統計的推測ないしは予測：推測統計
- 分析結果のプレゼンテーション：文書化，口頭発表
- 意思決定（終了もしくは最初に戻る）

10

PPDAC サイクル

- P : Problem
- P : Plan
- D : Data
- A : Analysis
- C : Conclusion

- CensusAtSchool



11

研究の種類

- 実験研究 (experimental study)
  - 処置効果の評価を意図. 実験条件の設定(無作為化など)が研究者自らの手でできる
- 観察研究 (observational study)
  - 処置効果の評価を意図. 観察条件の設定(無作為化など)が研究者自らの手でできない
- 調査 (survey)
  - 必ずしも処置効果の評価を意図しない.

- 前向き研究 (prospective study)
  - 条件を設定し, 時間を追って観測. コホート研究
- 後ろ向き研究 (retrospective study)
  - 現在の状態から過去にさかのぼって調査. ケース・コントロール研究

12

因果関係の確立

- Effect of Cause or Cause of Effect
- 統計学で主に扱うのは Effect of Cause

- ある処置 (treatment) に効果 (effect) があるか, あるとしたらどの程度か.
  - 新規開発医薬品, ICTを使った新しい教育方法, 新規の販売促進戦略, ある種の公共政策
- 一方で, Cause of Effect の探索も, 実用上重要
  - ある病気の原因は何か. どうやれば製品が売れるか. どうすれば学生の学力は上がるか.
  - 原因候補が特定できても, その次の段階として Effect of Cause の評価が必要

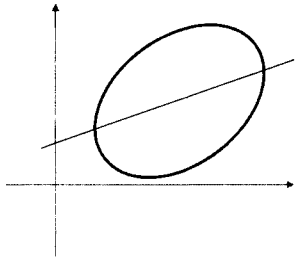
13

因果推論での登場物

- 目的:ある処置 (treatment) T の効果を，対照 (control) C との比較において評価
  - 「比較」は絶対に必要
  - 「薬を飲んだら病気が治った」，「WEBのデザインを変えたらページビューが増えた」だけでは不足
- 第一の変数:処置の割り付け変数:Z = 1 (T), = 0 (C)
- 第二の変数:結果変数:Y = 1（成功），= 0（失敗），あるいは連続量
- 第三の変数:（観測される）共変量:X（個体を特徴づけるもろもろの値で観測されるもの，通常は多数）
- 第四の変数:（観測されない）共変量:U（観測されないあらゆる要因）

14

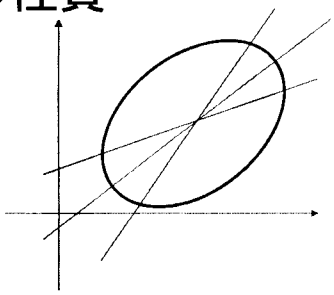
回帰モデル



- 単回帰モデル:$y = a + bx + e$
  - $y = ax + b + e$ ではない
  - $y = m_y + b(x - m_x) + e$
  - y：目的変数，x：説明変数，e：誤差項
  - a：定数項（通常は意味なし），b：回帰係数
  - m_y：y の平均，m_x：x の平均
  - 仮定:e は x とは独立に $N(0, \sigma^2)$ に従う
- 重回帰モデル:$y = b_0 + b_1x_1 + \cdots + b_px_p + e$
  - $y = m_y + b_1(x_1 - m_1) + \cdots + b_p(x_p - m_p) + e$
  - $y = b_0 + b_1x + b_2x^2 \cdots + b_px^p + e$
  - y：目的変数，$x_1, \ldots, x_p$：説明変数，e：誤差項
  - b_0：定数項，$b_1, \ldots, b_p$：（偏）回帰係数
  - 仮定:e は $x_1, \ldots, x_p$ とは独立に $N(0, \sigma^2)$ に従う

15

単回帰式 (y = a + bx) の性質



- 回帰直線は楕円の長軸ではない
- E[y | x] = a + bx : x を与えたときの y の条件付き期待値
  - x を定めたとき，対応する y は (a + bx) を中心にばらつく
- a : 定数項（通常は意味なし）
- b : x を1単位増加させたときの y の（平均的な）増分
- b の推定値 = Cov[x, y] / V[x]
  - V[x] = V[y] のときは b = ρ (= R[x, y], 相関係数)
- y から x への回帰式: x = c + dy
  - V[x] = V[y] のときは b = 1/ρ
- x が2値 (0 or 1) のときは b は各群の平均値の差

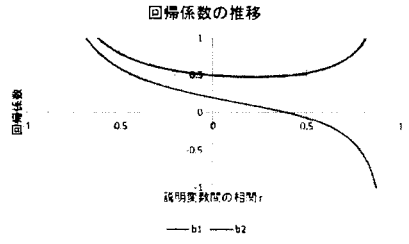
16

重回帰式 ($y = b_0 + b_1x_1 + b_2x_2$) の性質

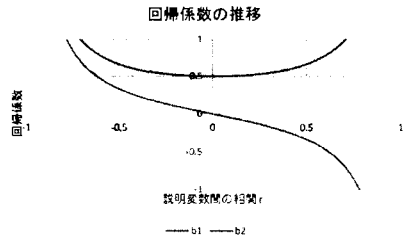
- $E[y \mid x_1, x_2] = b_0 + b_1x_1 + b_2x_2$: x_1, x_2 を与えたときの y の条件付き期待値
- b_1 の解釈1: x_2 の値を固定した上で，x_1 を1単位増加させたときの y の（平均的な）増分
- b_1 の解釈2: x_2 によって y のばらつきの説明をした残りの部分（y と x_2 との単回帰式の残差）に対し，x_1 を1単位増加させたときの y の（平均的な）増分
- 【重要】 b_1 の解釈は，$R[x_1, x_2] = 0$ であれば x_2 と無関係にできるが，$R[x_1, x_2] \neq 0$ のときは，x_2 に依存する
  - $R[x_1, x_2] = 0$ であれば，単回帰式 $y = a + bx$ と重回帰式 $y = b_0 + b_1x_1 + b_2x_2$ において，$b = b_1$ となる
- b_1 は x_2 に依存するので，x_2 として何をとるかが重要であり，b_1 の解釈をむやみに拡大してはならない

回帰係数の値の推移

- $r_1 = R[x_1, y] = 0.5$,
  $r_2 = R[x_2, y] = 0.2$
  と固定し, $r = R[x_1, x_2]$ を
  変化させたときの, 偏回
  帰係数 b_1, b_2 の動き



- $r_1 = R[x_1, y] = 0.5$,
  $r_2 = R[x_2, y] = 0$
  と固定し, $r = R[x_1, x_2]$ を
  変化させたときの, 偏回
  帰係数 b_1, b_2 の動き



回帰係数の計算式

- (x_1, x_2, y) の相関行列

$$\begin{array}{c} \\ x_1 \\ x_2 \\ y \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & y \end{array} \\ \begin{pmatrix} 1 & r & r_1 \\ r & 1 & r_2 \\ r_1 & r_2 & 1 \end{pmatrix} \end{array}$$

- b_1, b_2 の計算式(各分散は1に基準化)

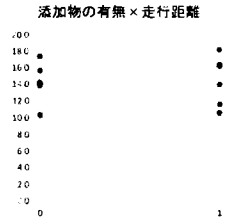
$$b_1 = \frac{r_1 - rr_2}{1 - r^2}, \quad b_2 = \frac{r_2 - rr_1}{1 - r^2}$$

- 説明変数間の相関 r が大きいと分母が小さくなって回帰係数が大きくなる.
- 相関 r が大きいと, $r_1 > 0$ であっても, 分子が負になることがある

添加物と走行距離の例 – 1

- 自動車のオイルにある添加物を入れることにより自動車の燃費(ガソリン1リットルあたりの走行距離)に差が出るかどうかを，添加物無では5台，添加物有では6台の自動車について，各走行距離を計測した．
- この添加物を加えることにより燃費が異なるかどうかを有意水準5%で両側検定せよ．
- 原因(処置):添加物の有無 (Z = 0, 1)
- 結果(効果):走行距離 (Y)
- 検定結果(2標本 t 検定):$t = -0.117$ ($P = 0.909$)

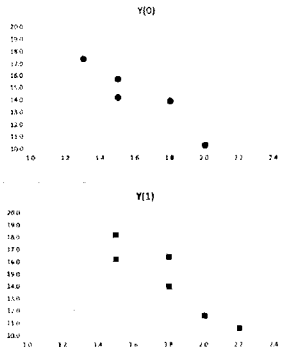
	添加物無	添加物有
ID	Y(0)	Y(1)
1	17.4	18.2
2	15.7	16.2
3	14.2	16.4
4	13.9	14.0
5	10.3	11.6
6		10.6

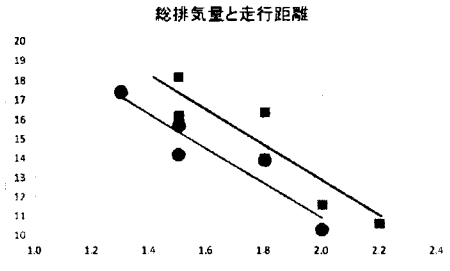


添加物と走行距離の例 – 2

- 原因(処置):添加物の有無 (Z = 0, 1)
- 結果(効果):走行距離 (Y)
- 共変量:自動車の総排気量(リットル)(X)

	添加物無		添加物有	
ID	X(0)	Y(0)	X(1)	Y(1)
1	1.3	17.4	1.5	18.2
2	1.5	15.7	1.5	16.2
3	1.5	14.2	1.8	16.4
4	1.8	13.9	1.8	14.0
5	2.0	10.3	2.0	11.6
6			2.2	10.6
平均	1.620	14.300	1.800	14.500
標準偏差	0.277	2.633	0.276	2.969





添加物と走行距離の例 – 3

- 共分散分析 (ANCOVA)
- モデル式: $Y = \alpha + \delta Z + \gamma X + \varepsilon$
- 効果量 (δ) の推定値: $d = 1.901$ ($P = 0.032$)

回帰統計	
重相関 R	0.925
重決定 R2	0.856
補正 R2	0.819
標準誤差	1.140
観測数	11

分散分析表

	自由度	変動	分散	分散比	有意 F
回帰	2	61.519	30.759	23.684	0.000
残差	8	10.390	1.299		
合計	10	71.909			

	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	29.605	2.283	12.965	0.000	24.340	34.871
Z	1.901	0.733	2.593	0.032	0.210	3.591
X	−9.448	1.374	−6.876	0.000	−12.616	−6.279

添加物と走行距離の例 – 4

- 共変量 (X) の値でマッチング: X の値が同じもののみをピックアップ
- 共変量の偏りを排除: 比較可能性を高める
- データ数が減少しているので統計的な有意性はないが，平均値の差に偏りはない

	添加物無		添加物有	
ID	X(0)	Y(0)	X(1)	Y(1)
1	1.3	17.4	1.5	18.2
2	1.5	15.7	1.5	16.2
3	1.5	14.2	1.8	16.4
4	1.8	13.9	1.8	14.0
5	2.0	10.3	2.0	11.6
6			2.2	10.6
平均	1.62	14.30	1.80	14.50
標準偏差	0.277	2.633	0.276	2.969

平均の差	0.20
t 値	−0.117
P 値	0.909

	添加物無		添加物有	
ID	X(0)	Y(0)	X(1)	Y(1)
1	1.5	15.7	1.5	18.2
2	1.5	14.2	1.5	16.2
3	1.8	13.9	1.8	16.4
4	2.0	10.3	2	11.6
平均	1.7	13.525	1.7	15.600
標準偏差	0.245	2.290	0.245	2.814

平均の差	2.075
t 値	−1.144
P 値	0.296

23

米国 SAT スコアの例 – 1

- 米国の SAT スコアは1980年に底を打ち，その後上昇に転じたとされる．
- 下の表は，人種別に見た平均 SAT スコアの推移
- White の平均は8点増加し，Non-Whiteの平均の増加は15点であるが，全体での平均の増加は7点

	平均スコア		
人種	1980	1984	差
White	924	932	8
Non-White	802	817	15
全体	890	897	7

cf. Wainer (1986)

24

米国 SAT スコアの例 – 2

- White の平均が 8 点増加，Non-White の平均が 15 点増加
- 全体の平均の増加は 7 点
- 足りない情報：受験者比率

$$924 \times 0.722 + 802 \times 0.278 = 890$$

$$932 \times 0.695 + 817 \times 0.305 = 897$$

- 第三の変数「受験者比率」の情報がないと解釈を誤る可能性

	平均スコア		
人種	1980	1984	差
White	924	932	8
Non-White	802	817	15
全体	890	897	7

	受験者比率		
人種	1980	1984	差
White	72.2	69.5	-2.7
Non-White	27.8	30.5	2.7
全体	100	100	0

喫煙の死亡率の例

- 喫煙習慣と死亡率について，カナダ，英国，米国の3つの調査研究が行われ，各喫煙習慣ごとの死亡率（1000人年）が報告された
- 調査時の平均年齢（第三の変数）を考慮しないと結論を誤る．
- 調整は，年齢階級ごとに求めた死亡率を融合

  cf. Cochran (1968)

Death rate	Study		
Smoking group	Canadian	British	U. S.
Non-smokers	20.2	11.3	13.5
Cigarettes only	20.5	14.1	13.5
Cigars, Pipes	35.5	20.7	17.4

Mean age	Study		
Smoking group	Canadian	British	U. S.
Non-smokers	54.9	49.1	57.0
Cigarettes only	50.5	49.8	53.2
Cigars, Pipes	65.9	55.7	59.7

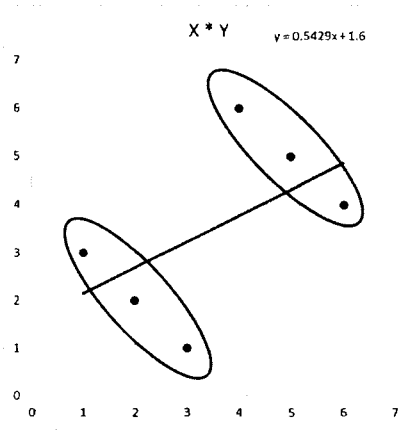
Adjusted D. R.	Study		
Smoking group	Canadian	British	U. S.
Non-smokers	20.2	11.3	13.5
Cigarettes only	29.5	14.8	21.2
Cigars, Pipes	19.8	11.0	13.7

さらに簡単な数値例

- 単回帰式：

  $y = 1.6 + 0.5429\,x$
- ダミー変数 d を入れた回帰式：

  $y = 4 + 6d - x$
- ダミー変数 d の導入により，各群での x と y との関係が正しく判断される
  - 添加物の例では，x の導入により d と y との関係が明らかになった



27

5つのべからず集

- 第三の変数を用いた調整法にはいくつかのものがある
  - マッチング，層別，共分散分析，. . .
- どうすればいいのか，に対する確固たる解答はないが，してはいけないことはある．

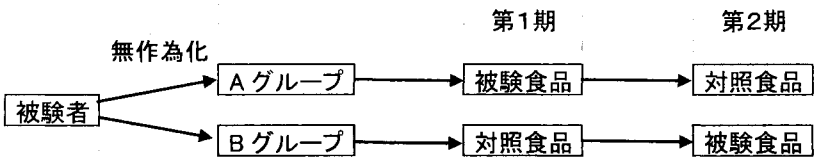
1. モデルを想定せず，やみくもに調整してはいけない
2. 処置に影響された変数を用いて調整してはいけない
3. モデルのチェックなしに外挿してはいけない
4. 調査対象とは異なる対象に関する変数で調整してはいけない
5. 調整したからといってその結果が常に妥当であると考えてはいけない

cf. Wainer (1989), Rosenbaum (1984)

28

クロスオーバー試験における層別

- 対照食品摂取後の値の高低（高群，低群）で2群に層別し，
  「効果量」＝「被験食品での結果」－「対照食品での結果」
  を計算
- その結果，高群での効果量に有意な差を認めた
- 「2．処置に影響された変数を用いて調整してはいけない」に抵触



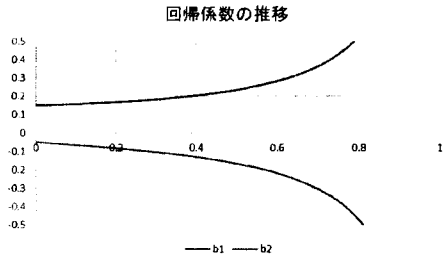
出生時体重と成人での血圧

- 出生時体重 (BW) が低いほど成人血圧 (BP) が高い（Barker 仮説）: BP = const + b_1BW において $b_1 < 0$
  - BP だけでなく，コレステロール値，心血管系イベントの発生率など
- 成人での BMI を説明変数に加える

$$BP = const + b_1BW + b_2BMI$$

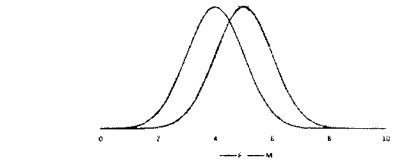
- b_1 の絶対値が大きくなる.
- r_1 = R[BW, BP] = – 0.05
- r_2 = R[BW, BMI] = 0.15
- $0 \leq r = R[BMI, BP] < 1$

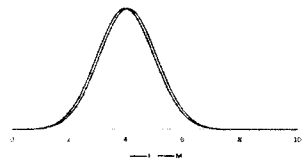
「2. 処置に影響された変数を用いて調整してはいけない」に抵触 cf. Tan, et al. (2005)

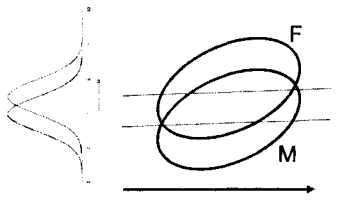


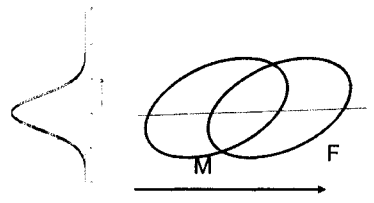
性差別？– 1

- 男性 (M) と女性 (F) で，賃金格差があるか
- 共変量 x = job performance



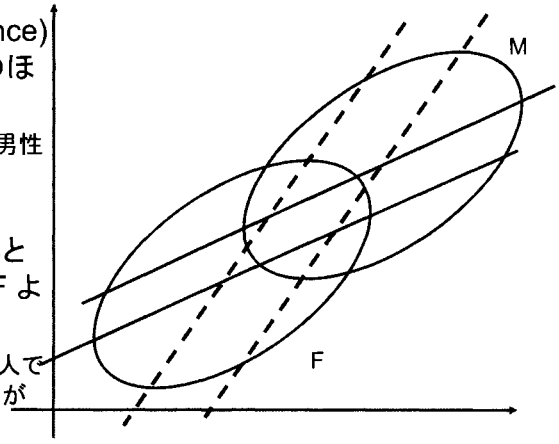






性差別？－2

- 同じ x (job performance) で見ると（実線），M のほうが F よりも大きい
  - 同じ成果であったとき，男性のほうが給料が高い
  - 女性に不利な差別
- 同じ y (salary) で見ると（破線），M のほうが F よりも大きい
  - 同じ給料をもらっている人で比較すると，男性のほうが成果が大きい
  - 男性に不利な差別



cf. Conway and Roberts (1983)

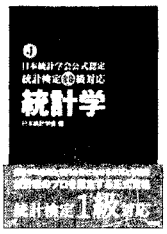
最後に：統計家はこう考える

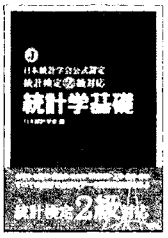
- 因果関係の確立には実験研究が gold standard
- 実験研究が必ずしも可能とは限らない
- 観察研究による因果推論では，実験研究に近づける努力
- 後ろ向き研究しかできないことも多い
  - 稀な事象の場合には，ほとんど唯一の方法論
- 現在そこにあるデータについては
- データの素性を明確に
- データ取得の 5W1H
  - Who, What, When, Where, Why + How
  - 統計では特に How が重要
  - Whom と How Much を加えて 6W2H (Wikipedia より)

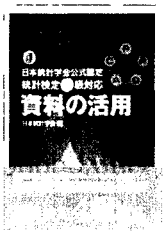
統計検定 (2014)

- 2014年11月30日(日)実施
  - 1級, 2級, 3級, 4級
  - 専門統計調査士, 統計調査士
  - 2級, 3級, 4級は年2回実施

- 学習マテリアルと問題集







参考文献(和書)

- 甘利俊一・狩野　裕・佐藤俊哉・松山　裕・竹内　啓・石黒真木夫 (2002) 多変量解析の展開　隠れた構造と因果を推理する. 岩波書店.
- Pearl, J.(著)黒木　学(訳) (2009) 統計的因果推論　モデル・推論・推測. 共立出版.
- 星野崇宏 (2009) 調査観察データの統計科学　因果推論・選択バイアス・データ融合. 岩波書店.
- 宮川雅巳 (2004) 統計的因果推論一回帰分析の新しい枠組み一. 朝倉書店.

参考文献(洋書－1)

- Berzuini, C., Dawid, P. and Bernardinelli, L. (eds.) (2012) *Causality. Statistical Perspectives and Applications*. John Wiley & Sons.
- Faries, D. E., Leon, A. C., Haro, J. M. and Obenchain, R. L. (Eds.) (2010) *Analysis of Observational Health Care Data Using SAS®*. SAS Institute.
- Morgan, S. L. (ed) (2013) *Handbook of Causal Analysis for Social Research*. Springer.
- Morgan, S. L. and Winship, C. (2007) *Counterfactuals and Causal Inference. Methods and Principles for Social Research*. Cambridge University Press.

参考文献(洋書－2)

- Rosenbaum, P. R. (2002) *Observational Studies, Second Edition*. Springer.
- Rosenbaum, P. R. (2010) *Design of Observational Studies*. Springer.
- Rothman, K. J., Greenland, S. and Lash, T. (2008) *Modern Epidemiology, Third Edition*. Wolters Kluwer.
- Rubin, D. B. (2006) *Matched Sampling for Causal Effects*. Cambridge University Press.
- Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002) *Experimental and Quasi-Experimental designs for Generalized Causal Inference*. Houghton Mifflin Company.

and others

参考文献(学術論文)

- Cochran, W. G. (1968) The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, **24**, 295-313.
- Conway, D. A. and Roberts, H. V. (1983) Reverse regression, fairness, and employment discrimination. *Journal of Business & Economic Statistics*, **1**, 75-85.
- Rosenbaum, P. R. (1984) The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A*, **147**, 656-666.
- Tu, Y.-K., West, R., Ellison, G. T. H. and Gilthorpe, M. S. (2005) Why evidence for the fetal origins of adult disease might be a statistical artifact: the "reversal paradox" for the relation between birth weight and blood pressure in later life (with discussion). *American Journal of Epidemiology*, **161**, 27-32.
- Wainer, H. (1986) Minority contributions to the SAT score turnaround: an example of Simpson's paradox. *Journal of Educational Statistics*, **11**, 239-244.
- Wainer, H. (1989) Eelworms, bullet holes, and Geraldine Ferraro: some problems with statistical adjustment and some solutions (with discussion). *Journal of Educational Statistics*, **14**, 121-199.

and many others

今後の活動予定

講演予定

- 統計関連学会連合大会(2014.9.13-16)
  - 於:東京大学(本郷キャンパス)
  - 9月13日午後チュートリアル講演「マッチングと統計解析」(3時間)
- 日本計算機統計学会シンポジウム (2014.11.14-15)
  - 於:沖縄科学技術大学院大学
- Kyoto International Conference on Modern Statistics (2014.11.17-18)
  - 於:京都国際会館

出版予定

- 岩崎 学 (2014 or 15) 統計的因果推論の基礎(仮題). 朝倉書店

iAnalysis

データサイエンスを活用した ビジネス拡大の事例

iAnalysis合同会社
代表・最高解析責任者　倉橋一成

設立経緯とビジョン

医療分野で培った最新の分析ノウハウをあらゆる場面に展開

iAnalysis最高解析責任者（CAO）が東京大学医学部と医療系コンサル会社で得たノウハウをもとに2011年に設立し、業務のシステム化、人々の行動変容、意思決定の手助け、社会貢献、人材教育などへさまざまな場面に展開しています。

"データイノベーションセンター"

「データは世界を変える」をビジョンとし、クライアント様の"**データイノベーションセンター**"となることを目指しています。

データ活用支援サービス提供先実績

2011年サービス開始から2014年の３年の間、

株式会社NTTドコモ
株式会社ベネッセコーポレーション
株式会社リクルートキャリア
株式会社インターネットイニシアティブジャパン
日本経済団体連合会
エーザイ株式会社
大鵬薬品工業株式会社
旭化成ファーマ株式会社
株式会社gumi
株式会社日経BP
株式会社ミクシィ
東京大学医学部付属病院

ほかには
大手自動車会社
大手携帯キャリア
大手製造会社
大手航空宇宙製造会社
Web広告ベンチャー
情報セキュリティベンチャーなど

など23業種、44社へサービス提供
（うち東証１部上場企業：36%）

設立者：最高解析責任者（CAO）
倉橋一成（Issei Kurahashi）

【経歴】
東京大学医学部博士課程卒 (2011)
統計家、コンサルタント、データサイエンティスト、健康学博士
Statistician , Consultant, Data Scientist, Ph.D

【専門/スキル】
統計解析、コンサルティング、データサイエンス、医療データ分析
R, SAS, SPSS, Python

【バックグラウンド】
・2005：NPO日本臨床研究支援ユニット、解析担当
・2007、2009：スタットコム株式会社、統計解析者
・2009～2010：帝京大学、医師への統計コンサルタント
・2010：キャピタルメディカ株式会社、プロジェクトメンバー

・2011：iAnalysis合同会社 設立
・2011～2012：東大病院　特任助教（兼務）
・2013～　　：東大病院　特任研究員
・2013～　　：経団連　　医療ビッグデータ研究会　委員

ブログ閲覧数100万回：http://d.hatena.ne.jp/isseing333/
国際論文10本以上：http://isseing.jimdo.com/%E7%B5%8C%E6%AD%B4/

2014/3/27 読売新聞朝刊

iAnalysisのサービス

ビジネスセクター向け

1．ビジネス拡大のための データ活用支援サービス

アカデミックセクター向け

2．論文レビューサービス

また、ご要望に応じてセミナーや講演も行います。
詳しくは別途資料にてご案内致します。

ホームページ：ianalysis.jp
問い合わせ先：contact@ianalysis.jp
お電話：03-6868-3490

事業拡大のためのデータ活用支援サービス

情報通信技術（IT）
Internet Technology
＋
ビジネスインテリジェンス
Business Intelligence
＋
統計解析
Statistical Data Analysis
＝
新しいアイディア
Innovative Idea

弊社はデータ活用のノウハウによって、様々なアイディアを生み出します。

弊社のアイディアはパートナー様の実行力に掛け合わせることで、ビジネス価値を何倍にも拡大します。

分析
↓
アイディア
×
実行力
＝
価値

新たな経営指標の開発により
年間約120億円の売上増加を発見

東証1部上場　大手自動車メーカー

レセプトデータの集計分析によって
年間1億円の売上増加を発見

病床数200床　地方中核総合病院

※サービス料金等、内容の詳細は別途資料にてご案内致します。

論文レビューサービス

これまでインパクトファクター(IF)の高い
国際雑誌に多数投稿してきたノウハウを活かし、
統計解析方法のレビューやアドバイスを行います。

【これまでの投稿実績】
J Bone Miner Metab(2.2), Clin Exp Nephrol(1.2),
Kidney Int(7.9), Am J Hypertens(3.8),
Diabetes Care(8.1), Insect Science(1.7),
Nephrol Dial Transplant(3.3), Clinical Cancer Res(7.8),
PLOS ONE(4.0)

【これまでのサービス提供実績】
東京大学医学部付属病院、東京大学工学部、
大鵬薬品工業株式会社、近畿大学、福島医科大学

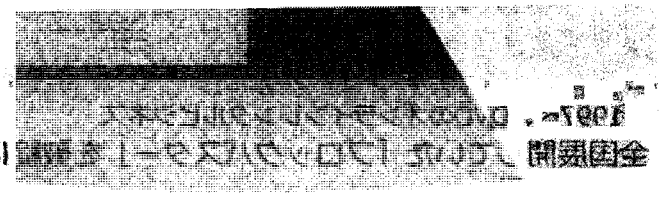
インパクトファクターとは
論文の引用数を指標化したもので、一般的に2～3くらいから国際論文として認められています。
科学誌で最高峰と言われるNatureが36、Scienceが31です。

IFが2以上の雑誌は無制限にレビュー
- メールにてやりとりを行う
- 国内論文やIFが2未満の雑誌への投稿は、年間3報まで
- 最初の1報は無料

弊社担当者を共著または謝辞に入れてもらう

年間契約で30万円（税抜）、研究室単位での契約

データサイエンスとは



iAnalysis

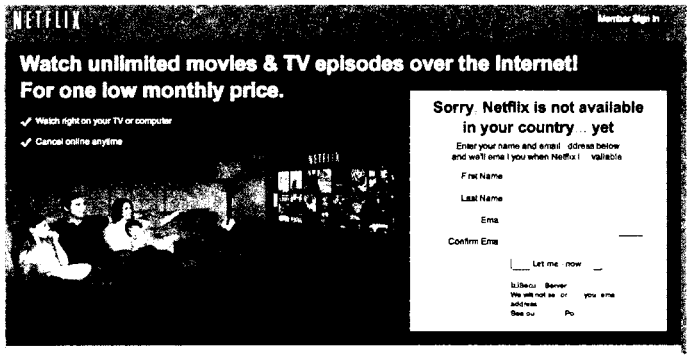
身近なサービスで行われている分析・データサイエンス

iAnalysis

9

具体的にどんな「分析」が行われているか

- ネットフリックス



1997～、DVDのオンラインレンタルビジネス

- 全国展開していた「ブロックバスター」を破綻に追いやるほど成長

iAnalysis

10

レコメンドエンジン：シネマッチ

- 顧客の好みを分析して映画をレコメンド（推奨）する
  - アマゾンのレコメンドが有名

配送方法：スロットリング（throttling）

- たまにしか借りない、利益率の高い顧客に優先してDVDを届けたいが、頻繁に借りる会員が「不公平だ」と感じてしまう（→訴訟も）
- 「利益の最適化」と「公平な配送」のバランスを計算するアルゴリズムを開発

頒布（はんぷ）権の購入金額

- 新しい映画の頒布権を購入するとき、過去に借りられた「似ているジャンルの映画」と同じくらいレンタルされるだろう

A/Bテスト

- 新しいサービスを作るとき、それが本当に効果があるかどうか、A/Bテストによって常にチェックする

分析の活用事例

分析事例

「分析力のある企業」の成功事例
- GOOGLE：リスティング広告
- Amazon：商品のレコメンデーション
- PayPal：不正検知
- キャピタルワン：クレジットカードのパーソナライズ
- ネットフリックス：ビデオのレコメンデーション

一般事例
- ダイレクトマーケティングの効果アップ
- ユーザーの離反防止
- 株式投資自動化

iAnalysis事例
- 経営企画の仮説検証、論文研究のための仮説検証
- Web訪問者の属性予測、広告効果の高いユーザーセグメントの発見
- 婚活サイトのユーザー分析
- 化粧品会社の顧客分析
- 新しいレコメンデーションシステム企画立案のための調査データ分析
- 情報の不正流出検知アルゴリズムの開発

鉄鋼製造会社（Rocky Mountain Steel Mills）

背景

2005年に価格競争のためシームレス鋼管製造を打ち切ったが、原油価格が高騰したために原油採掘会社からの需要が高まった。

課題

鋼管製造の再開を検討。
しかし意思決定のためのコスト分析の信頼性が低いと感じていた。

分析

プロフィット・インサイトという分析ソフトを導入し、工場を再稼働させるべきかどうか分析結果をみながら毎月検討。
12月に損益分岐点を超え、さらに予測モデルによってその後も価格上昇が見込まれる状況になって初めて、製造の再開を行った。

成果

早期に生産再開した場合の損失4300万ドルを回避

クレジット会社（キャピタル・ワン）

背景

1990年代、「情報ベース戦略」を打ち立てる。
「まだ顔を見たことない2億の人達について情報を集め、集めた情報を基にして、長期的な作戦を練る」

分析

データベースの整備、分析などを精緻に行うことで、「高額の商品をあっさりクレジットで買い、長期にわたってゆっくり返済する客」が最も優良顧客であることが判明。

成果

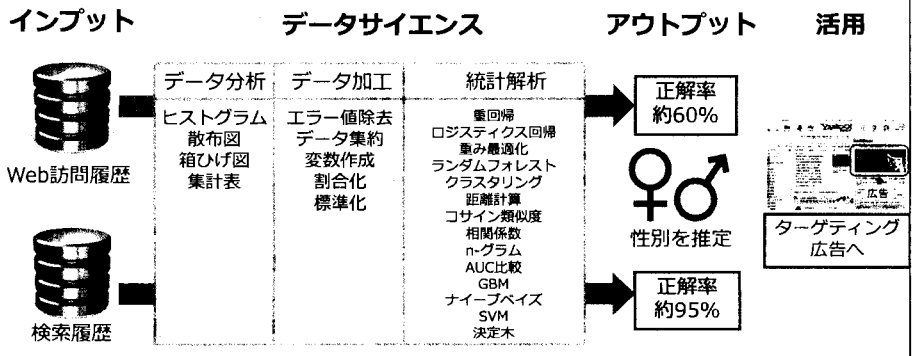
業界で初めて「リボルビング機能」をカードに搭載し、新商品開発につながった。

現在では1日に300回のマーケティング調査。
譲渡性預金の利息、ロールオーバーのための優遇措置、最低必要残高などと、顧客定着率との間にはっきりとした関係があることが判明。
→定着率の87%アップ、新規顧客開拓コストの83%ダウン

iAnalysis

15

事例：インターネットサービス関連企業
Web訪問履歴を利用したリバースプロファイリング

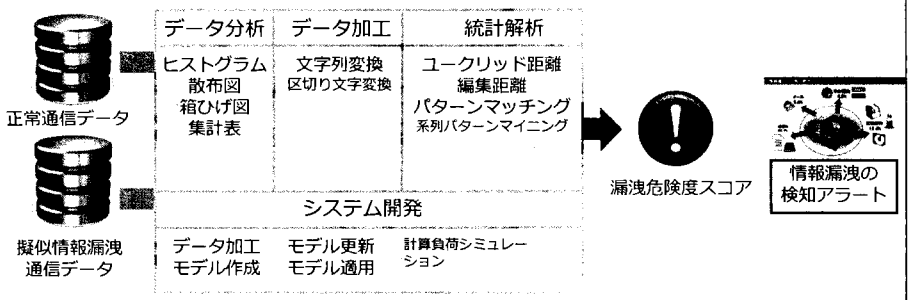


プロジェクトメンバー：データサイエンティスト×2（弊社）
マネージャー、DBエンジニア（クライアント）
プロジェクト期間：３ヶ月（他の分析も並行）

アーキテクチャ

16

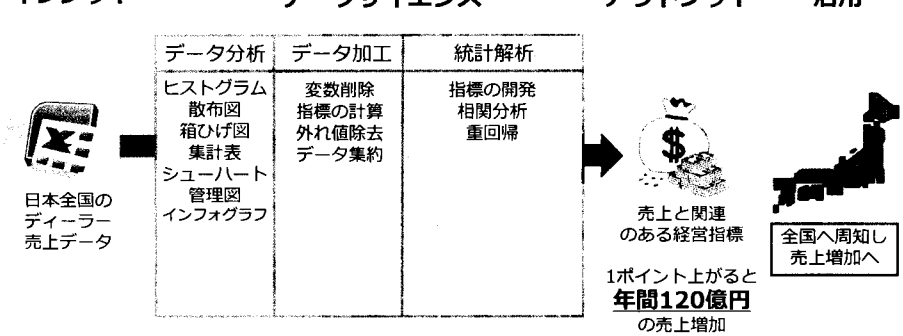
事例：業界トップ セキュリティ関連企業
社内PCからの情報漏洩防止アルゴリズムの開発



プロジェクトメンバー：データサイエンティスト×2（弊社）
マネージャー、エンジニア（クライアント）
プロジェクト期間：1.5ヶ月

アーキテクチャ

事例：大手自動車メーカー 本社企画部
経営企画のための新しい切り口の経営指標を開発



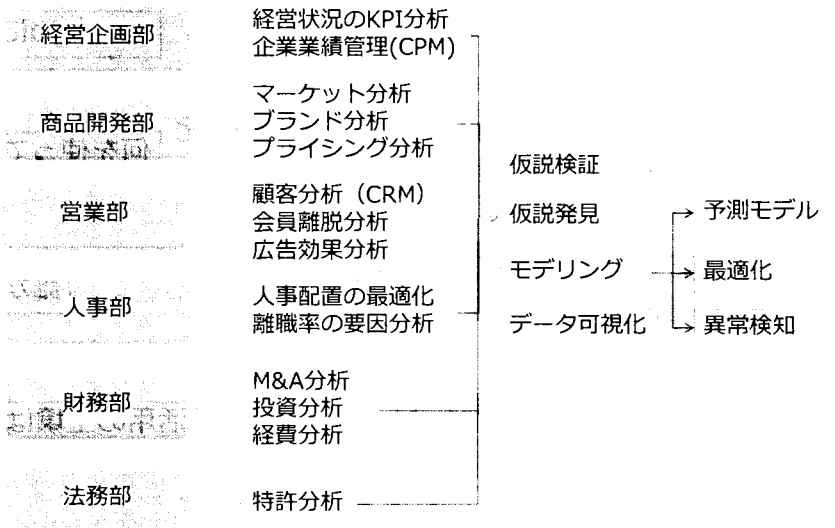
プロジェクトメンバー：データサイエンティスト×2（弊社）
企画部課長、技術者
プロジェクト期間：3ヶ月

アーキテクチャ

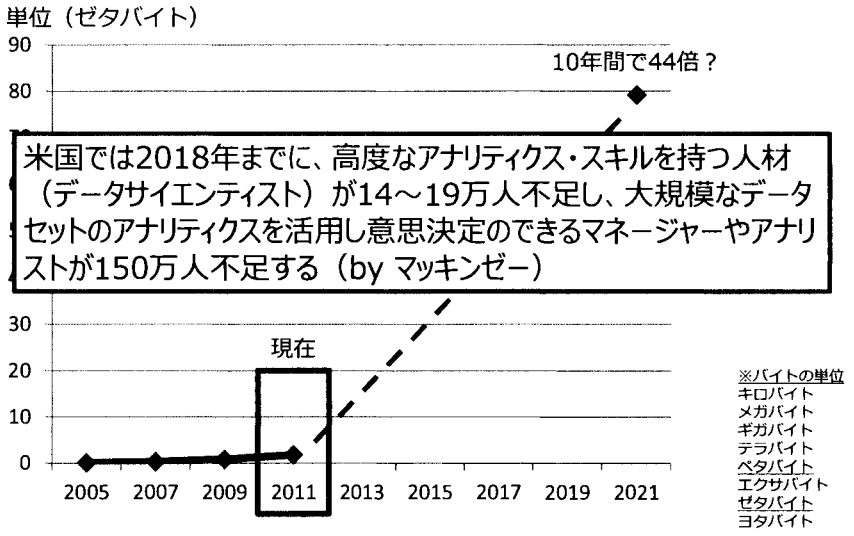
そもそも分析の用途・目的は？

- ①仮説検証、現状分析
  ⇒マーケティング系
  - 市場データの分析、広告効果の分析
  - ECサイトの分析
- ②仮説発見
  - 自然言語の共起ネットワーク
  - 企業間の取引ネットワーク
- ③最適化、異常検知、予測モデル
  ⇒システム、研究開発系
  - レコメンド
  - 在庫、仕入れ、配送の最適化
  - 不正アクセス解析
  - リスティング広告最適化
  - 癌の予後予測
  - 入院日数予測

部署ごとでの効果的な分析手法



データが溜まるスピードに分析が追いつかない



米国では2018年までに、高度なアナリティクス・スキルを持つ人材（データサイエンティスト）が14～19万人不足し、大規模なデータセットのアナリティクスを活用し意思決定のできるマネージャーやアナリストが150万人不足する（by マッキンゼー）

社内でデータを活用する際の課題

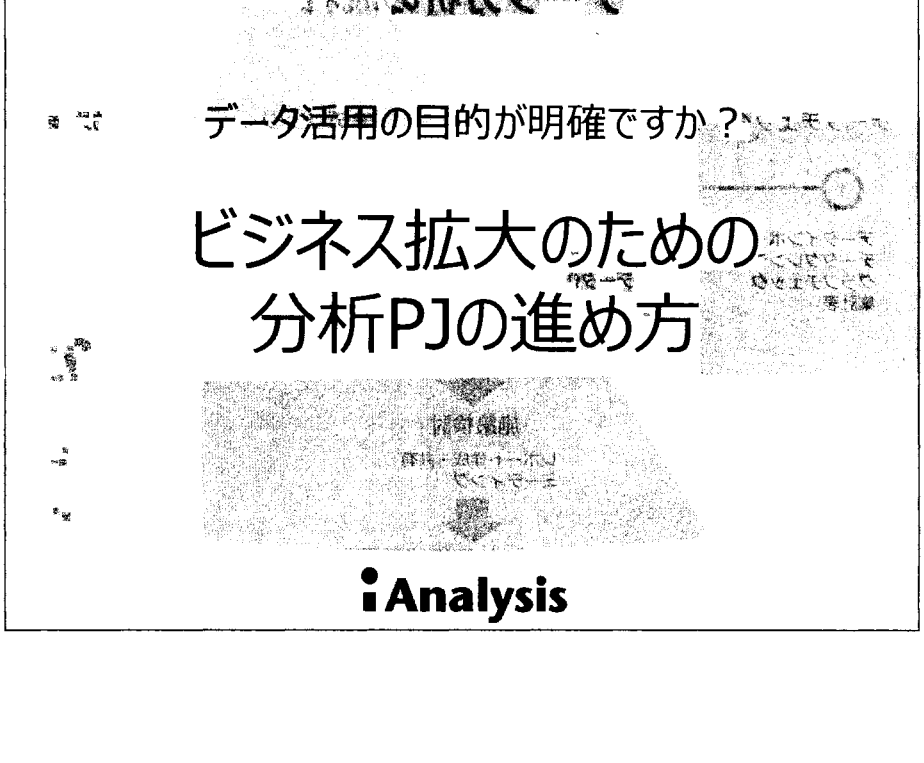
- データ活用の目的が明確ですか？ 何のために？
- 目的に合った分析ツールやシステムを適切に選べていますか？ 何を使って？
- これまでデータ分析を活用してきた社員が何人いますか？ 誰が？
- これまでの勘や経験だけで経営判断しませんか？ 活用の土壌は？

データ活用の目的が明確ですか？

ビジネス拡大のための分析PJの進め方

iAnalysis

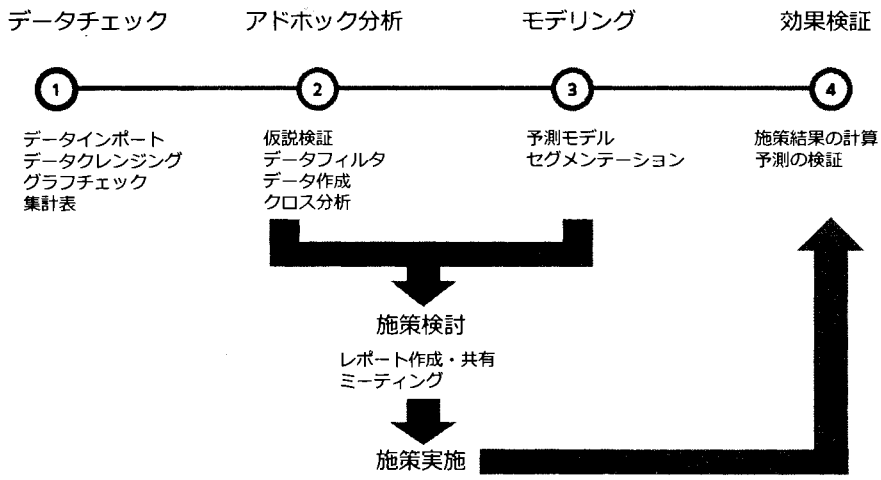
ビジネス拡大のための分析PJの進め方



iAnalysis

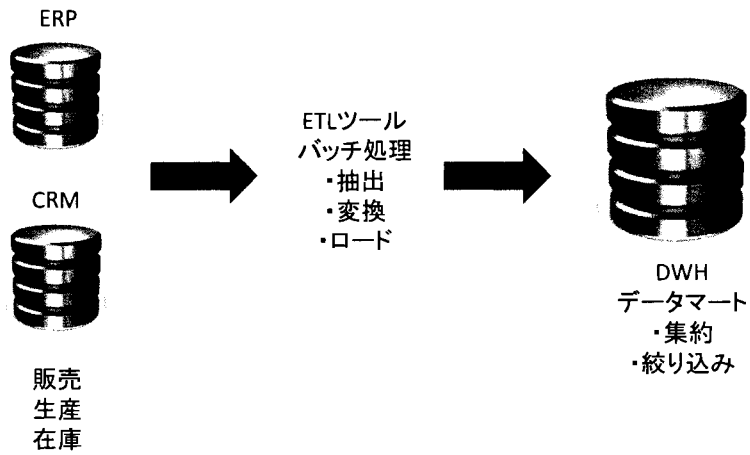
24

データ分析の流れ



Phase I　データチェック

- データの収集や加工



データクレンジングの課題

※分析の中でデータクレンジングに費やす時間の割合は70～90%※

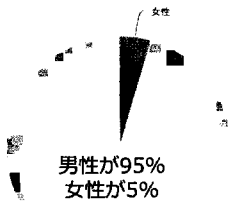
全国の健診データを分析するプロジェクト

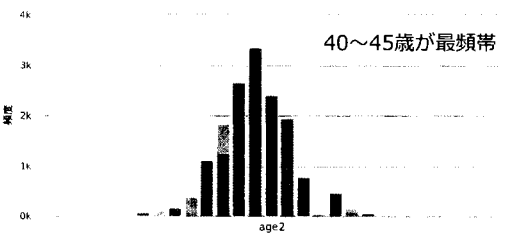
- 概要
  - 目的：特定健診データを収集しクレアチニン測定の意義を分析する
  - 全国数十の市町村からデータ収集
  - 約60万人
  - 5年間は追跡目標
- データクレンジングが最大のネック
  - 国保によってcsvファイルの仕様が微妙に違う
    - 尿蛋白などが1~6になっていたり-, +-, …, +++になっていたり
  - 入力ミス、エラー値がある

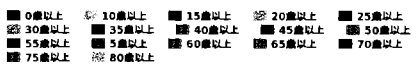
課題

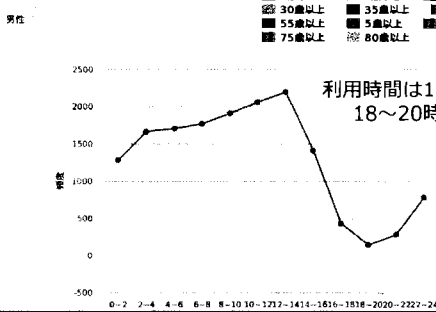
- 巨大データをどうやってクレンジングするか？
  - データを全て可視化することができない
  - ロジックを組んだからといってコンピュータに任せっきりは危険
  - 「データが分かる人」が逐次モニターする必要がある

Phase II　アドホック分析



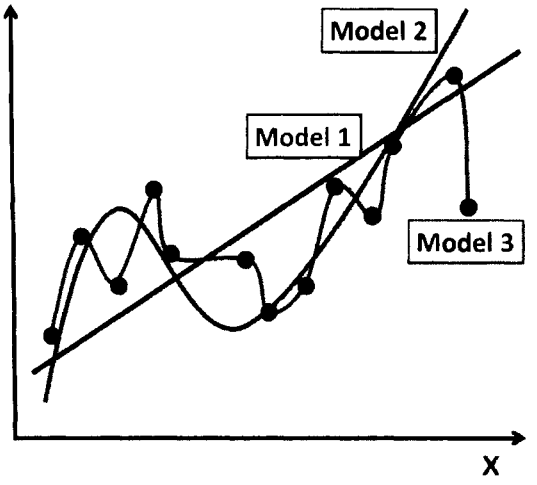






Phase III　モデリング

- データに様々な「モデル」を当てはめて、情報を探索する



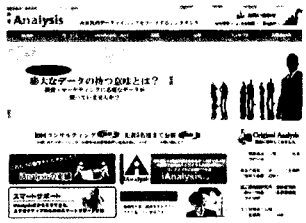
iAnalysis

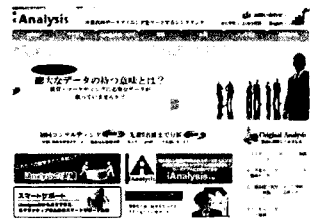
29

Phase IV　効果検証（A/Bテスト）

1000人

1000人





コンバージョン、直帰率、サイト滞在率、リピート率などを比較

iAnalysis

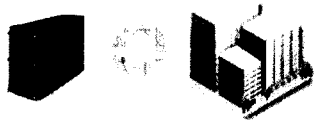
30

目的に合った分析ツールやシステムを
適切に選べていますか？

インフラの整備、目的に合った ツール・システム導入

iAnalysis

様々な分析サービス



インフラ・DB　　ETL・DWH　　ソフトウェア BAツール　　ソリューション BIツール

ORACLE	PENTAHO	SAS	R	SPSS	Salesforce
Hadoop	Neteeza	Lavastorm		JMP	GoogleAnalytics
MySQL	Greenplum			Mathmatica	GoogleAdwords
PostogreSQL				Statistica	
Amazon RedShift				Stata	CRMサービス
				エクセル	ERPサービス

iAnalysis

ビッグデータのデータベース

- 様々なところに記録されているデータを統合する
  - 社内の部署連携
  - データベースエンジニア、インフラエンジニア
- 大規模データを扱う必要
  - 数100GB～数10TB
  - Facebookは1日に約100TBのデータが発生
  - Googleは約200億（？）のサイトから検索を行っている（約400TB？）
  - Amazonは数千万アイテムの中からリコメンド（推奨）している
- 「分散処理」によって高速に処理を行う
  - Hadoop（ハドゥープ）
    - Googleの基盤技術であるMapReduceをJavaでオープンソース実装した分散処理のフレームワーク

iAnalysis

データマイニングとデータサイエンスの違い

- データマイニング
  - （大量の）データから有益な情報を掘り起こす（マイニング）こと
  - 分析対象のデータは「排気データ」なことが多い
  - 技術的な視点が強い
- データサイエンス
  - データを適切に分析することで、正しい意思決定を行う
  - 目的、仮説を持って意識的にデータを溜めて分析しよう
  - 技術を何のためにどう活かすかという視点が強い
- Google：「次の10年で熱い職業は統計学」
  - あらゆるデータが記録される時代
  - データをどのように有効活用するか！
  - http://www.publickey1.jp/blog/10/10_3.html
- Facebook：「データサイエンティスト」を公募

iAnalysis

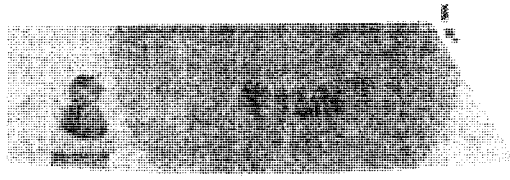
分析＋ビジネス業務Webサービスを開発中

- データチェックから効果検証をワンストップで
- 分析内容を手軽にレポーティング
- ビジネス業務機能を便利に使えて、データも溜めれる
- RedShiftに接続してビッグデータも扱える
- 豊富なテンプレートにより高度な分析もワンクリックで

iAnalysis

これまでデータ分析を活用してきた社員が
何人いますか？

人材確保、人材育成



iAnalysis

データサイエンティストってどんな人？

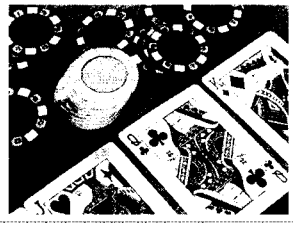
①統計学を駆使して②データ分析することで
③ビジネスインパクトのある結果を産み出す人

- それぞれの企業内で活躍している人
  - Google：広告効果を分析
  - Amazon：レコメンドエンジンの効果を分析
  - DeNA：マーケティング分析
  - リクルート：多種サービスの分析
- 外から企業に入って活躍している人
  - 富士通、NEC、日立：様々な企業のデータ分析
  - アクセンチュア、野村総研：様々な企業へコンサルティング
  - iAnalysis：様々な企業へコンサルティング

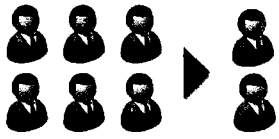
データサイエンティストのスキルセット・人物像

データ分析は統計学（Statistics）が基盤

- 統計学：「経験的に得られたデータを分析し法則性を見出す学問」
- 政治・ギャンブルなどのニーズから生まれた



サンプリング調査



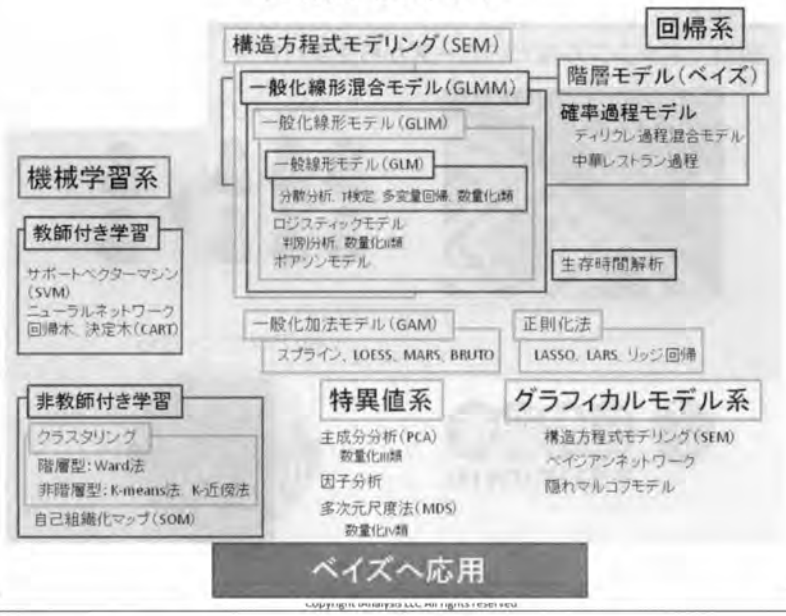
webアクセス・広告

iAnalysis

回帰系：利用する場面とモデル、手法の全体像

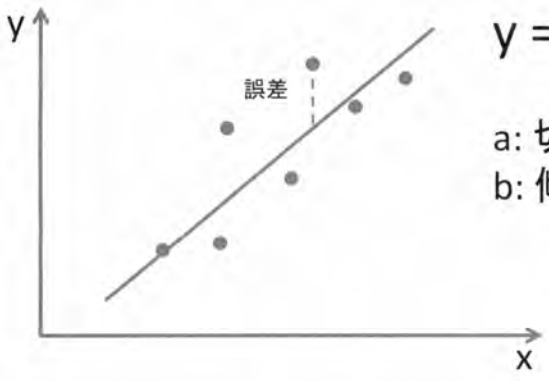
結果変数: y	説明変数: x	手法
連続値	2値	t検定
	3つ以上のカテゴリー	分散分析
	連続	線形単回帰、線形重回帰
	カテゴリー、連続	共分散分析
2値	カテゴリー	分割表、ロジスティック回帰
	連続	ロジステック回帰など
	カテゴリー、連続	ロジステック回帰
3つ以上のカテゴリー	カテゴリー	分割表
	カテゴリー、連続	名義ロジステック回帰
順序	カテゴリー、連続	順序ロジステック回帰
カウント値	カテゴリー	対数線形モデル
	カテゴリー、連続	ポアソン回帰
生存時間	カテゴリー、連続	Cox回帰
相関のある値、グループ値	カテゴリー、連続	混合効果モデル

分析手法マップ



回帰モデルのイメージ

- 結果変数yが連続値の場合



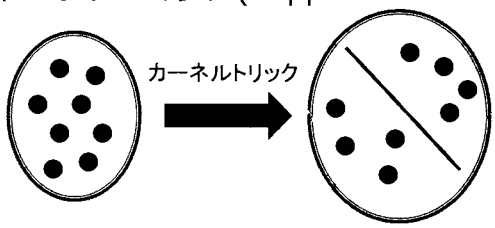
$$y = a + bx$$

a: 切片
b: 傾き

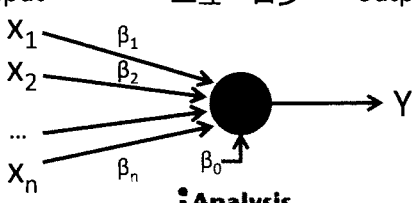
iAnalysis

機械学習

- サポートベクターマシン(Support Vector Machine; SVM)



- ニューラルネットワーク（Neural Network; NN）



iAnalysis

43

手法ごとの主な用途

線形回帰（重回帰、ロジスティック回帰、ポアソン回帰、Cox回帰等）	要因分析、予測・記述モデル
決定木	要因分析、予測・記述モデル、要因の可視化、ルール抽出
一般化加法モデル	時系列プロットの可視化、スムージング
混合効果モデル、階層ベイズ	クラスター効果の調整して正確な効果を確認
主成分分析、因子分析	変数のグルーピング、変数縮約
LASSO	特徴量選択
教師付き機械学習（SVM、ランダムフォレスト、GBM等）	予測
非教師付き機械学習（クラスタリング、SOM等）	データをグループ分けし、グループの特徴を確認
グラフィカルモデリング（SEM、ベイジアンネットワーク等）	要因分析、要因の可視化
検定	群間差が統計的に有意かどうか確認

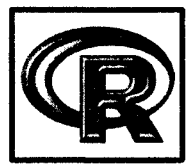
iAnalysis

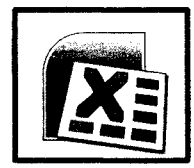
44

分析のためのエンジニアリングスキル

C++

d3.js





UNIX

可視化、レポーティング系

基本は抑えておき、
必要となったらその場で
調べて利用するという
スタンス

データベース系

スクリプト、分析系

データサイエンティストを目指すなら知っておきたいRパッケージ10個＋α

- randomForest：超強力な汎用予測モデル
- RPostgreSQL, RMYSQL, RMongo, RODBC, RSQLite：各種データベースへの接続
- plyr：データ集約
- reshape2：データ加工
- forecast：時系列予測
- (stringr：文字列操作)
- (lubridate：日付操作)
- (sqldf：SQLライクなデータ操作)
- (ggplot2：綺麗なプロットを描く)
- qcc：品質管理

- party：決定木が綺麗に描ける
- gbm：randomForestより汎用性の高い超強力な予測モデル
- survival：生存分析
- caTools, Epi：予測モデルの性能評価に必要なROC曲線が描ける、AUCを計算できる
- XLConnect：エクセルのデータを読み込める、Rオブジェクトをエクセルに保存できる

TokyoR

- @yokkunsが主催
- 月1回のペース、現在25回開催
- Rを中心にして、さまざまな分析トピックスを発表

- 発表者は毎回5～10人
- 参加者は各社の分析屋＋学生
- 参加人数は40～80人
- 参加方法：TwitterやGooglegroupで告知されるのでATNDで登録
  - https://groups.google.com/group/r-study-tokyo?hl=ja&pli=1

- 会場：Nifty社のセミナールーム（北新宿）
- 資料置き場：
  http://lab.sakaue.info/wiki.cgi/JapanR2010?page=%CA%D9%B6%AF%B2%F1%C8%AF%C9%BD%C6%E2%CD%C6%B0%EC%CD%F7

iAnalysis

フォローすべき分析系Twitterアカウント

- isseing333
- gepuro
- milionsmile
- teramonagi
- yokkuns
- teikaw
- horihiro
- dichika
- wdkz
- hamadakoichi
- doryokujin
- sas20yen
- bob3bob3
- mikedewar
- kan_yukiko
- rindai87
- a_bicky
- mihiog
- kos59125
- smly
- iakiyama
- iAnalysisLLC
- hoxo_m
- AntiBayes

- hiroue_harada
- mikado_hito
- Hiro_macchan
- wakuteka
- phosphor_m
- Rbloggers
- triadsou
- Rbloggers
- _kohta
- fuzzysphere
- Door_intoSummer
- langstat
- aad34210
- tyatsuta
- sfchaos
- shima__shima
- sakaue
- yanaoki
- NPHard
- StatsGuild
- spssjapan

分析関連ブログ、web上の情報

- R advent calenderで書かれた25ブログ
  - http://atnd.org/events/22039
- Rに関する情報まとめ
  - Rを使えるようになるための10のこと＋α
  - http://d.hatena.ne.jp/isseing333/20110917/1316231082
- Facebookページ
  - iAnalysis: http://www.facebook.com/ianalysis
  - StatsGuild: http://www.facebook.com/StatsGuild.jp
  - データマイニング: https://www.facebook.com/pages/%E3%83%87%E3%83%BC%E3%82%BF%E3%83%9E%E3%82%A4%E3%83%8B%E3%83%B3%E3%82%B0/148756641848693
  - WebMining: http://www.facebook.com/WebMiningStudies
- ポータルサイト
  - KDnuggets: http://www.kdnuggets.com/
- Data Science Central： http://www.datasciencecentral.com/

KAGGLE（http://www.kaggle.com/）

- データサイエンスハッカソン@ロンドン
  - 2012年7月21日
- 医療データによる入院日数予測
  - 1位には2.4億円（2013年4月3日締め切り）
- 信用スコアの改善
- レコメンデーションシステム
- サッカーワールドカップ優勝国の予測
- 高速道路の渋滞予測
- …
- 現在48イベント
  - http://www.kaggle.com/competitions

iAnalysisセミナー

学問	統計学	統計学の基礎、回帰
	機械学習	教師付き機械学習、非教師付き機械学習
エンジニアリング	SQL	select where, join
	スクリプト	Python, unix, サーバー操作
	R	Rの基本操作、データ操作、基本関数、回帰、 パッケージを使った応用

参考書

iAnalysis

分析系で読んでおくべき本

- 統計学
  - 統計学入門
  - 自然科学の統計学
  - 多変量解析入門
  - Elemental of Statistical Learning（修士以上レベル）
  - Data Mining for Decision Making
- エンジニアリング
  - Rによるやさしい統計学
  - 統計学：Rを用いた入門書
  - Rによるデータサイエンス
  - データサイエンティスト養成読本
- ビジネス、事例
  - 分析力を武器とする企業
  - 分析力を駆使する企業

iAnalysis

これまでの勘や経験だけで
経営判断しませんか？

企業内の分析文化

iAnalysis

なぜデータ分析が重要か？

- 勘や経験や度胸　vs データ
- 製品やサービスの価格を決めるとき、過去に類似の商品が類似の状況でいくらなら売れたのかというデータを無視して勘で決めたら？
- 人材を採用するとき、そのポストではどんなスキルや適性が高業績につながるか、過去のデータを分析せずに採用担当者の直感で決めたら？
- 在庫水準をデータに基づく最適水準に維持せず、「このくらいがちょうどいい」という漠然とした経験で決めたら？

iAnalysis

企業の分析力の発展の五段階

1. 分析力に劣る企業
2. 分析力の活用が限定的な企業
3. 分析力の組織的な強化に取り組む企業
4. 分析力はあるが決定打に至らない企業
5. 分析力を武器とする企業

ステージ	組織戦略		人			インフラ
	目標	現状	スキル	経営陣のコミットメント	企業文化	
1. 分析力に劣る企業	顧客・市場・競合について知る。	分析はほとんど行われていない。	なし	なし	データアレルギー。直感に頼る。	データがない。精度が低い。定義が曖昧。システムがばらばら。
2. 分析力の活用が限定的な企業	データ分析の経験を自主的に蓄積し、トップの関心を引く。	ごく狭い範囲でしかデータ収集・分析が行われていない。	一部の部門にアナリストがいるが孤立している。	特定事業や戦術的な対応に限られている。	客観的なデータが必要と感じている。一部の部門では関心が高まっている。	各事業ばらばらにデータを収集している。重要なデータが欠落している。システムが統合されていない。
3. 分析力の組織的な強化に取り組む企業	組織横断型でデータ収集・分析を行う。全社共通の業績評価指標を設定する。データ分析で事業機会を探す。	分析プロセスは各部門不統一である。	多くの部門にアナリストがいるが、ネットワーク化されていない。	分析力を競争優位にすることに一部の幹部が興味をもち始めた。	経営陣は事実を重んじる姿勢を打ち出しているが、抵抗に遭っている。	システムやソフトウェアは整い、データウェアハウスも拡張中。

ステージ	組織戦略		人			インフラ
	目標	現状	スキル	経営陣のコミットメント	企業文化	
4. 分析力はあるが決定打に至らない企業	組織横断型の分析プラットフォームを構築し、組織として分析力を身につける。	データ分析がある程度まで業務プロセスに組み込まれている。	スキル開発は行われているが、まだ水準に達していない、または適材適所ではない。	経営陣のサポートが得られている。	事実に基づく意思決定の浸透を図っている。	データの精度は高く、全社的な分析戦略もある。分析環境は整っている。
5. 分析力を武器とする企業	データ分析から多くの隠されていた事実を導き出す。継続的にデータやシステムの改善を図る。	データ分析が定着し、高度に統合化されている。	高度なスキルを備え、意欲のある専門家がそろっている。周辺業務はアウトソースされている。	CEOを筆頭に経営陣が積極的に取り組んでいる。	事実に基づいて意思決定を下す。実験し学習する姿勢が浸透している。	組織横断型のシステムが整備・運用されている。

iAnalysis

57

効率良く分析を進めるための組織編成

集中管理モデル

経営チーム

分析グループ　部門A　部門B

機能モデル

経営チーム

部門A　部門B

分析グループ

分散配置モデル（非推奨）

経営チーム

部門A　部門B

分析グループ　分析グループ

iAnalysis

58

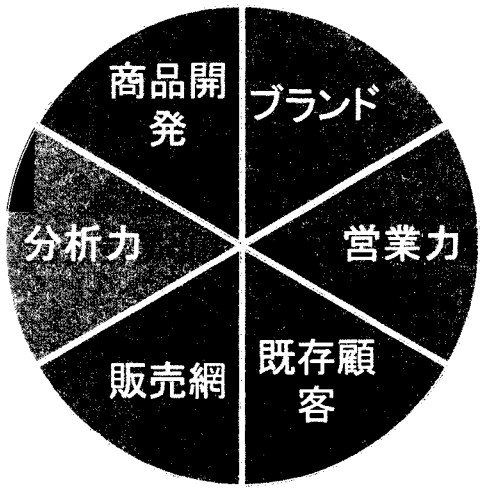
データサイエンス業界の課題を解決するには

- データ活用の目的が明確ですか？ 何のために？
  ⇒目的を持って効果的に進める、アウトソーシングも活用
- 目的に合った分析ツールやシステムを適切に選べていますか？ 何を使って？
  ⇒多種多様のツールがある中で適切に選択する
- これまでデータ分析を活用してきた社員が何人いますか？ 誰が？
  ⇒セミナー・研修による教育、専門家のコンサルティングを活用
- これまでの勘や経験だけで経営判断しませんか？ 活用の土壌は？
  ⇒勘だけじゃなくて数字も見るようにする

分析が適さない状況は？

- 時間がないとき
  - 瞬時に決定する必要がある
- 前例がないとき
  - 小規模なA/Bテスト
- 過去の事例が当てにならないとき
  - 株価
  - 地震
- 意思決定者がきわめて経験豊富なとき
  - データ収集と分析プロセスを頭の中で行なってしまう
- 変数が計測できないとき

分析力で競合他社と差をつける



iAnalysis

61

”データイノベーションセンター”

「データは世界を変える」をビジョンとし、
クライアント様の“データイノベーションセンター”
となることを目指しています。

contact@ianalysis.jp
ianalysis.jp

iAnalysis

62

チュートリアル

LOGISTICプロシジャによる解析と最新の機能拡張

映画「タイタニック」のロマンティック回帰

浜田知久馬
東京理科大学

Analysis of SAS LOGISTIC procedure and new functions.

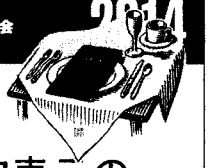
Chikuma Hamada
Tokyo University of Science

要旨：

生存・死亡のような2値データの多変量解析を行うLOGISTICプロシジャについて，モデル構築の方法のチュートリアルを行う．

またLOGISTICプロシジャのV.9.3までの機能拡張について紹介する．

キーワード：LOGISTIC ROC曲線，多重性調整オッズ比，Firth's Penalized Likelihood



内容

1）オッズ比,予測確率のプロット等，グラフ表示のODS GRAPHICSの機能の充実

2）CONTRAST，ESTIMATE，LSMEANS，LSMESTIMATE，ODDSRATIOO文を利用することで
共変量と多重性の双方を同時に調整した解析

3）ROC曲線を，ROC文で作成し，ROCCONTRAST文で複数のモデル間でAUCの比較

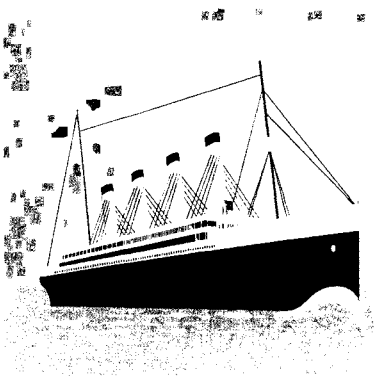
4）MODEL文のFIRTHオプションでFirth's Penalized Likelihoodに基づいた推測

3

自治体、大学によるエコシステムの実証 SASユーザー総会 2014

映画「タイタニック」

ジャック　ローズ



運命を変える恋がある

ロジステイック回帰で
ロマンテイック回帰
ジャックとローズの
ラブロマンスが
かなう確率を分析

4

SASによるロジステイック回帰

LOGISTICプロシジャ

CATMODプロシジャ

GENMODプロシジャ

INSIGHTプロシジャ

PROBITプロシジャ

GLIMMIXプロシジャ

何が影響をあたえるか

ロジスティックモデル

ある現象の発生する確率（割合）$p(\mathbf{x})$を

その現象の生起を説明するために観測された変数群　$\mathbf{x} = (x_1, x_2, \cdots, x_r)$

で説明するモデルを考える，r個の変数群$\mathbf{x}$の下で現象が生起するという条件付き確率を　$p(\mathbf{x})$　で表し，これを，

$$p(\mathbf{x}) = \Pr\{\text{生起} \mid x_1, x_2, \cdots, x_r\} = F(x_1, \cdots, x_r)$$

という分布関数Fを用いてモデル化

ロジスティックモデル

分布関数Fを用いてモデル化:

$$
\begin{aligned}
p(\mathbf{x}) &= \Pr\{\text{生起} \mid x_1, x_2, \cdots, x_r\} \\
&= F(x_1, x_2, \cdots, x_r) \\
&= \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r)}
\end{aligned}
$$

ロジスティック関数:

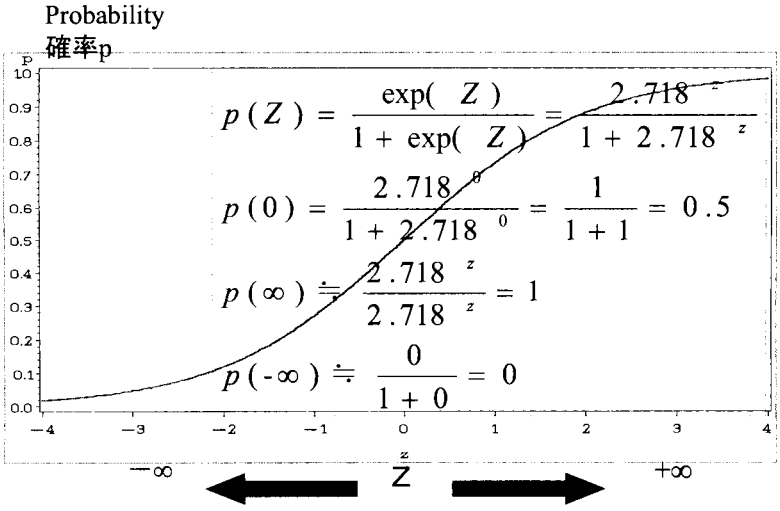
$$Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r$$

$$F(Z) = \frac{\exp(Z)}{1+\exp(Z)} = \frac{1}{1+\exp(-Z)} = \frac{2.718^z}{1+2.718^z}$$

7

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

ロジスティック曲線



$$p(Z) = \frac{\exp(Z)}{1+\exp(Z)} = \frac{2.718^z}{1+2.718^z}$$

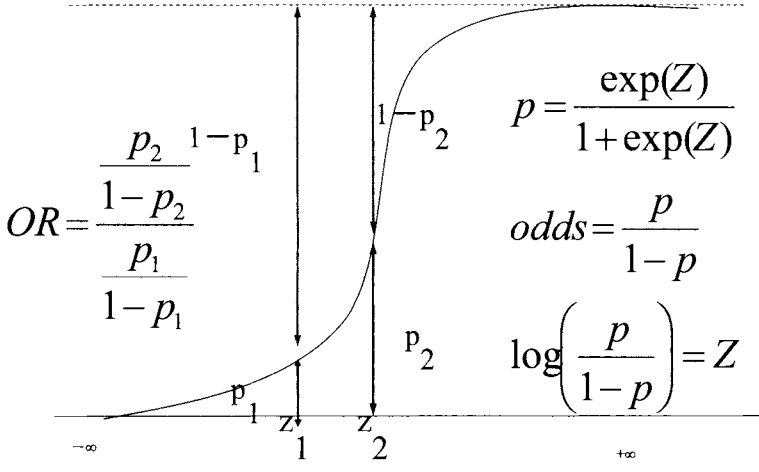
$$p(0) = \frac{2.718^0}{1+2.718^0} = \frac{1}{1+1} = 0.5$$

$$p(\infty) \fallingdotseq \frac{2.718^z}{2.718^z} = 1$$

$$p(-\infty) \fallingdotseq \frac{0}{1+0} = 0$$

ロジスティック曲線とオッズ

イベント発現確率p



$$OR = \frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}}$$

$$p = \frac{\exp(Z)}{1+\exp(Z)}$$

$$odds = \frac{p}{1-p}$$

$$\log\left(\frac{p}{1-p}\right) = Z$$

ロジット(logit)

$$p(\mathbf{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r)}$$

$$= \frac{1}{1 + \exp(-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_r x_r)}$$

ここから，式変形を行うと

$$\log \frac{p(\mathbf{x})}{1 - p(\mathbf{x})} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r$$

上記左辺を$p(\mathbf{x})$のロジット(対数オッズ)という

オッズ(odds):p/(1-p)

$$\log\frac{p(\mathbf{x})}{1-p(\mathbf{x})}=\beta_0+\beta_1(x_1+1)+\beta_2x_2+\cdots+\beta_rx_r$$

$$\frac{p(\mathbf{x})}{1-p(\mathbf{x})}=\exp\left(\beta_0+\beta_1(x_1+1)+\beta_2x_2+\cdots+\beta_rx_r\right)$$

$$=\exp\left(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_rx_r\right)\times\boxed{\exp(\beta_1)}$$

上記左辺を$p(\mathbf{x})$の$1-p(\mathbf{x})$に対するオッズという

$\exp(\beta)$:他の変数固定で，x を 1 単位変化させたときのオッズ比

説明変数が1つの場合

	薬剤 −	薬剤 +
イベント +	5	10
イベント −	95	90

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1x$$

$$p=\frac{\exp\left(\beta_0+\beta_1x\right)}{1+\exp\left(\beta_0+\beta_1x\right)}$$

$$e^{\beta_1}=OR=\frac{10\cdot95}{5\cdot90}=2.11$$

x=0:drug− x=1:drug+

イベント +	5	10
イベント −	95	90

likelihood（尤度）

尤度（L）＝モデルの下でデータが得られる確率

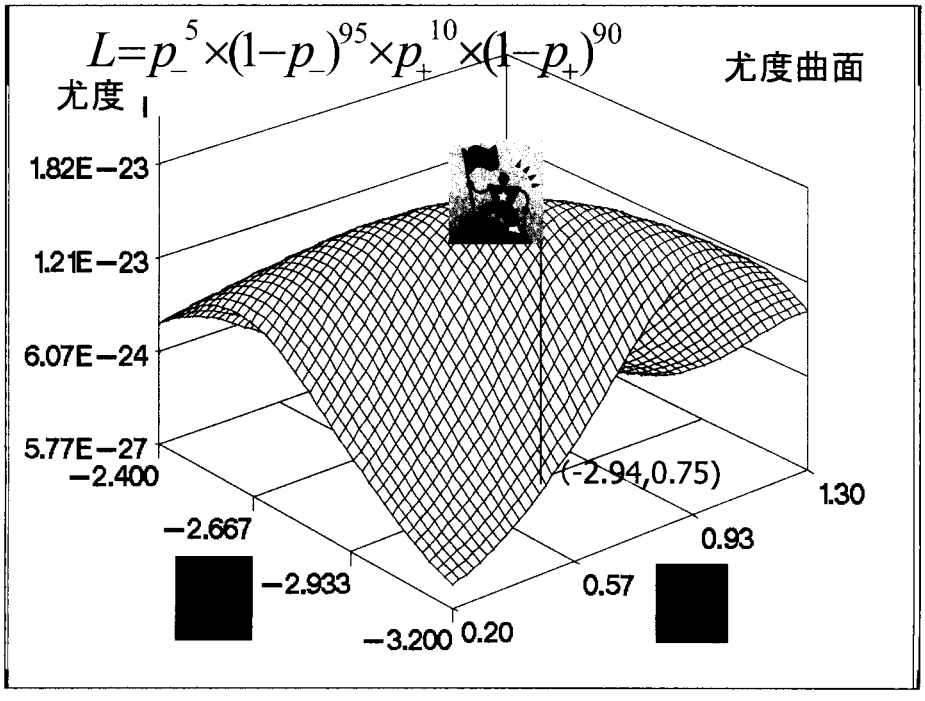
$$L = p_-^{5} \times (1-p_-)^{95} \times p_+^{10} \times (1-p_+)^{90}$$

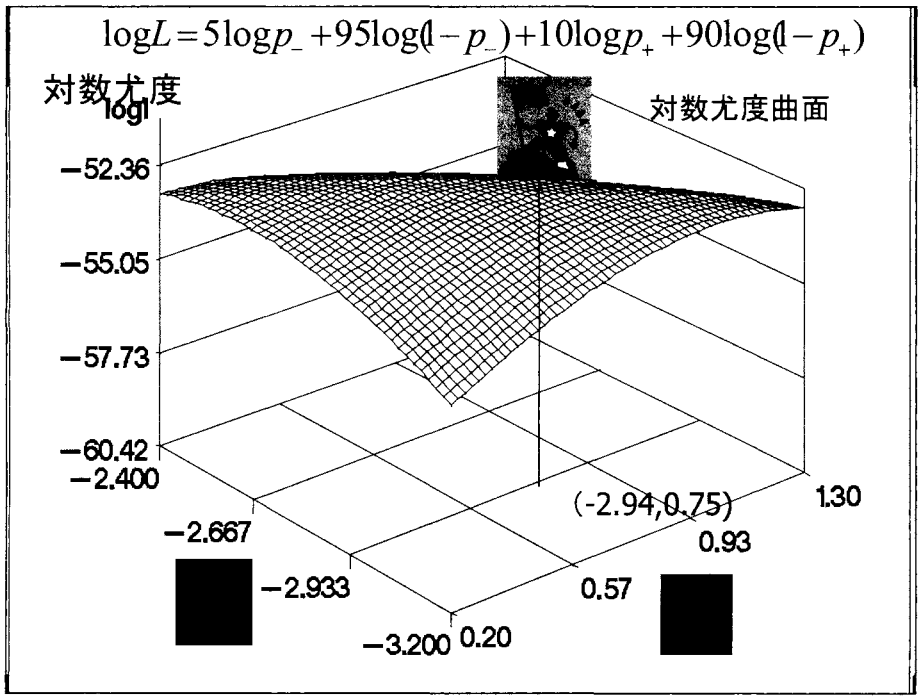
$$p_- = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}, \quad p_+ = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}$$

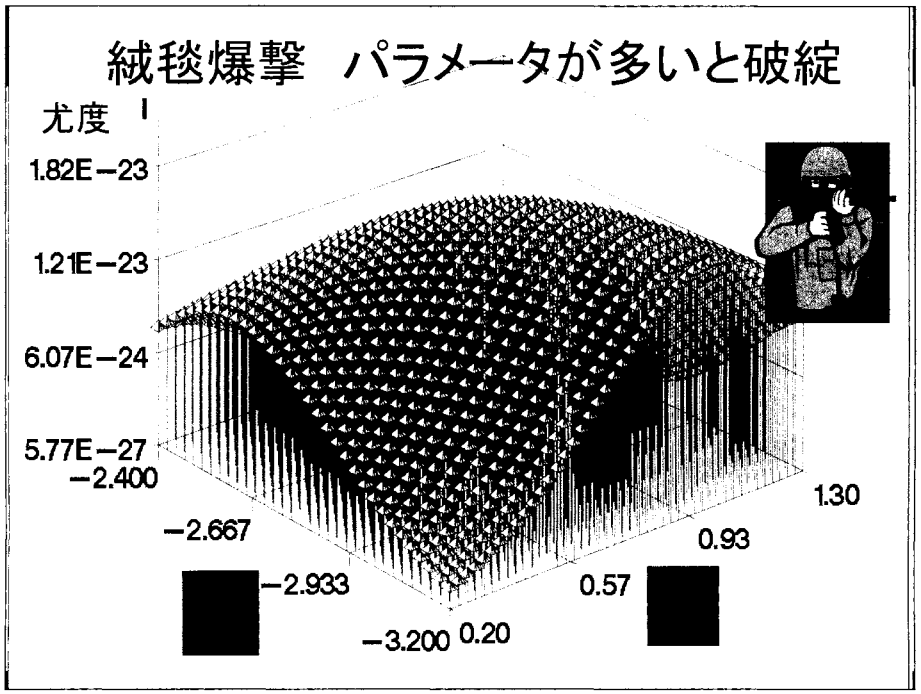
最尤法：β_0, β_1の値を動かしてLが最も大きくなるようにする方法

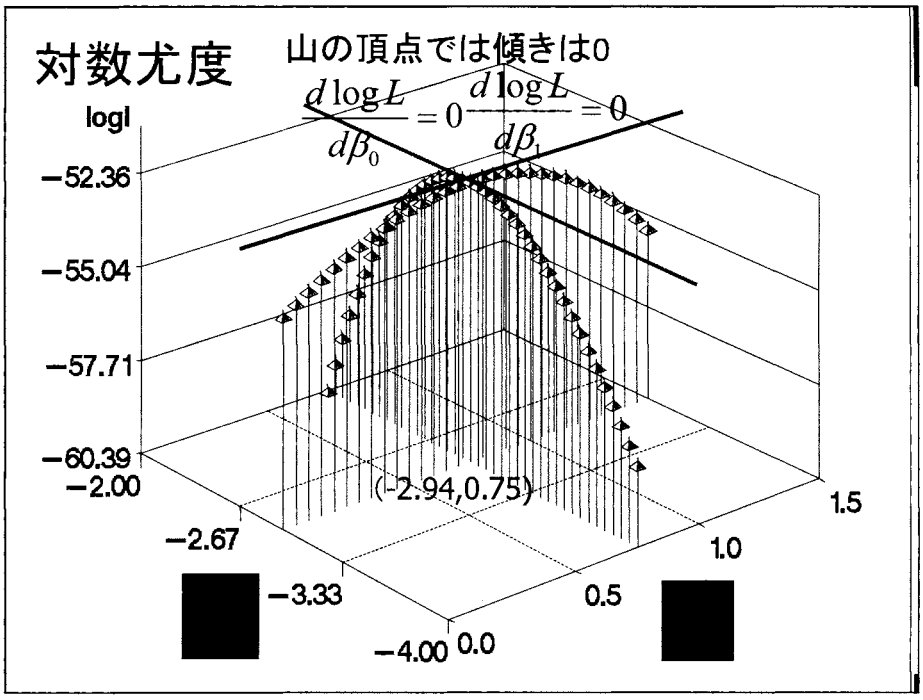
MLE：Maximum Likelihood Estimator

13




$\log L = 5\log p_- + 95\log(1-p_-) + 10\log p_+ + 90\log(1-p_+)$
対数尤度
logl
対数尤度曲面
−52.36
−55.05
−57.73
−60.42
−2.400
−2.667
−2.933
−3.200
0.20
0.57
0.93
1.30
(-2.94,0.75)


絨毯爆撃　パラメータが多いと破綻
尤度
1.82E−23
1.21E−23
6.07E−24
5.77E−27
−2.400
−2.667
−2.933
−3.200
0.20
0.57
0.93
1.30



イベント +	a	c
イベント −	b	d

対数尤度とスコア関数

$$L = p_-^{\,a} \times (1-p_-)^b \times p_+^{\,c} \times (1-p_+)^d$$

$$\log L = a\log p_- + b\log(1-p_-) + c\log p_+ + d\log(1-p_+)$$

$$U(\beta_1) = \frac{d\log L}{d\beta_1} = c-(c+d)p_+ = 0 \Rightarrow \boxed{p_+ = \frac{c}{c+d}}$$

$$U(\beta_0) = \frac{d\log L}{d\beta_0} = a+c-(a+b)p_- -(c+d)p_+ = 0 \Rightarrow \boxed{p_- = \frac{a}{a+b}}$$

最尤推定量

イベント+	a	c
イベント−	b	d

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x$$

$$\hat{\beta}_0 = \log \frac{p_-}{1-p_-} = \log \frac{a/(a+b)}{1-a/(a+b)} = \boxed{\log \frac{a}{b}}$$

$$\hat{\beta}_0 + \hat{\beta}_1 = \log \frac{p_+}{1-p_+} = \log \frac{c/(c+d)}{1-c/(c+d)} = \log \frac{c}{d}$$

$$\hat{\beta}_1 = \log \frac{p_+}{1-p_+} - \log \frac{p_-}{1-p_-} = \boxed{\log \left(\frac{bc}{ad} \right)}$$

19

データセットNeuralgia(N=60)

Example 53.2 Logistic Modeling with Categorical Predictors

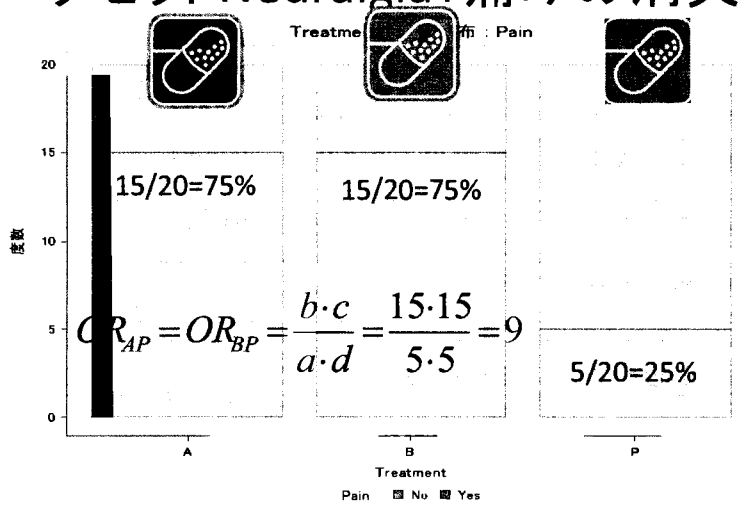
高齢者のNeuralgia(神経痛)に対する
鎮痛薬の試験
反応変数:Pain(痛みの有無:Yes, No)の消失率
群変数:Treatment(治療:P(プラセボ), A, B)
共変量:Sex(性別:F, M)
Age(治療開始時の年齢:連続量)
Duration(履病期間:連続量(月))

20

データセットNeuralgia(N=60)

```
Data Neuralgia;
   input Treatment $ Sex $ Age Duration Pain $ @@;
   datalines;
P F 68  1 No  B M 74 16 No  P F 67 30 No  P M 66 26 Yes
B F 67 28 No  B F 77 16 No  A F 71 12 No  B F 72 50 No
B F 76  9 Yes A M 71 17 Yes A F 63 27 No  A F 69 18 Yes
B F 66 12 No  A M 62 42 No  P F 64  1 Yes A F 64 17 No
P M 74  4 No  A F 72 25 No  P M 70  1 Yes B M 66 19 No
B M 59 29 No  A F 64 30 No  A M 70 28 No  A M 69  1 No
B F 78  1 No  P M 83  1 Yes B F 69 42 No  B M 75 30 Yes
P M 77 29 Yes P F 79 20 Yes A M 70 12 No  A F 69 12 No
B F 65 14 No  B M 70  1 No  B M 67 23 No  A M 76 25 Yes
P M 78 12 Yes B M 77  1 Yes B F 69 24 No  P M 66  4 Yes
P F 65 29 No  P M 60 26 Yes A M 78 15 Yes B M 75 21 Yes
A F 67 11 No  P F 72 27 No  P F 70 13 Yes A M 75  6 Yes
B F 65  7 No  P F 68 27 Yes P M 68 11 Yes P M 67 17 Yes
B M 70 22 No  A M 65 15 No  P F 67  1 Yes A M 67 10 No
P F 72 11 Yes A F 74  1 No  B M 80 21 Yes A F 69  3 No
```


データセットNeuralgia：痛みの消失率



LOGISTICの基本プログラム

```
ods graphics on;
proc logistic PLOTS=(ODDSRATIO EFFECT)
    data=neuralgia;
    class treatment sex/param=glm;
    model pain= treatment;
    oddsratio treatment;
    run;
ods graphics off;
```


パラメータ推定値とオッズ比

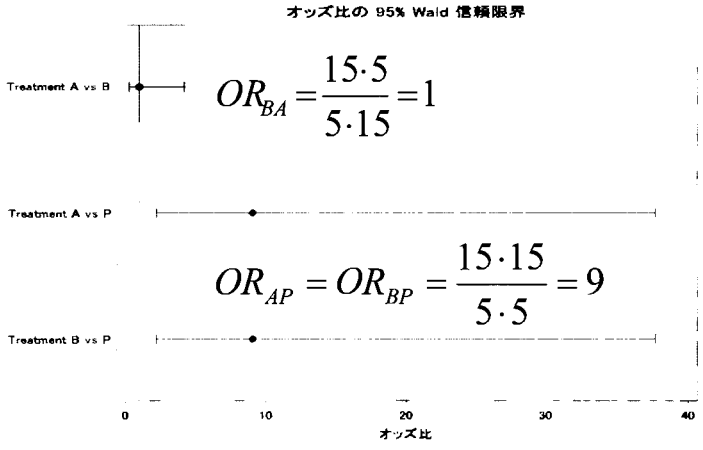
最尤推定値の分析

パラメータ		自由度	推定値	標準誤差	Wald カイ 2 乗	Pr > ChiSq
Intercept		1	-1.0986	0.5164	4.5261	0.0334
Treatment	A	1	2.1972	0.7303	9.0521	0.0026
Treatment	B	1	2.1972	0.7303	9.0521	0.0026
Treatment	P	0	0	.	.	

オッズ比推定と Wald による信頼区間　　oddsratio treatment

ラベル	推定値	95% 信頼限界	
Treatment A vs B	1.000	0.239	4.184
Treatment A vs P	9.000	2.151	37.659
Treatment B vs P	9.000	2.151	37.659

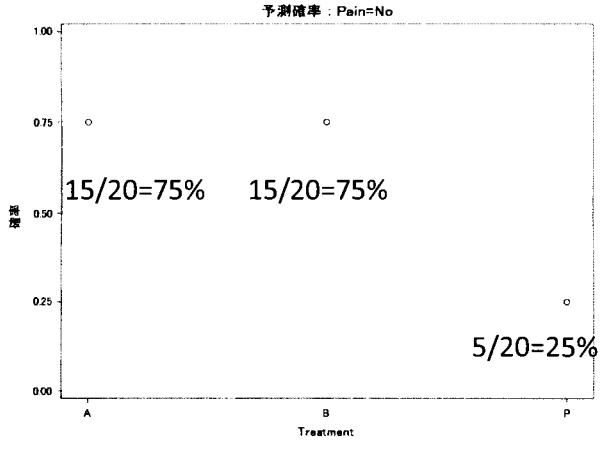
PLOTS=(ODDSRATIO)
デフォルトはdiff=all



$$OR_{BA} = \frac{15 \cdot 5}{5 \cdot 15} = 1$$

$$OR_{AP} = OR_{BP} = \frac{15 \cdot 15}{5 \cdot 5} = 9$$

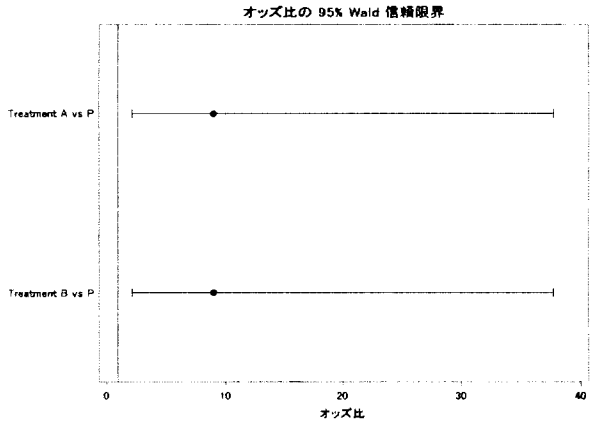
PLOTS=(EFFECT)



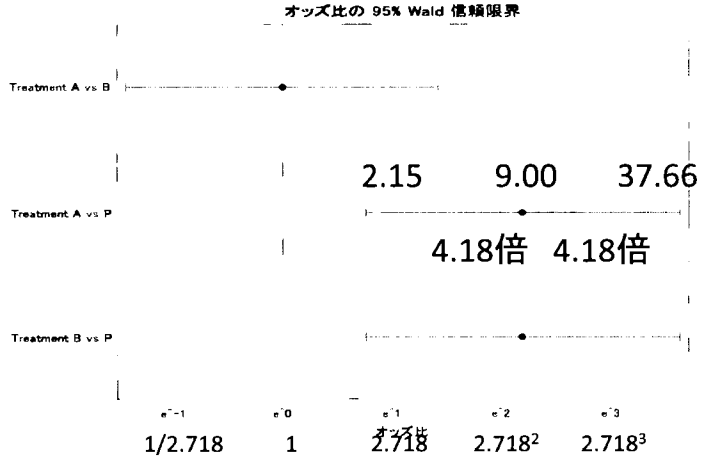
oddsratio treatment /diff=ref;

基準群との比較

デフォルトはdiff=all

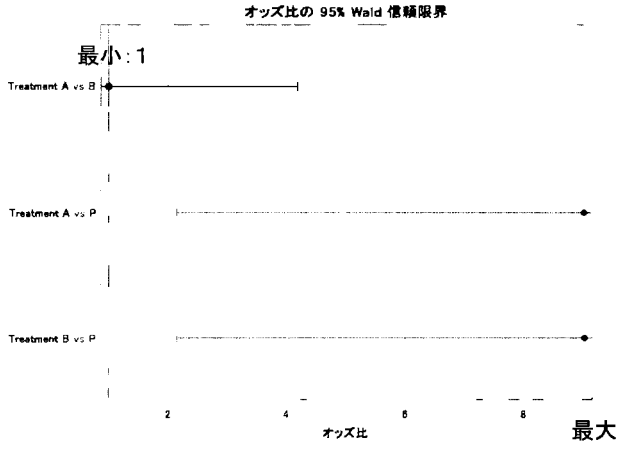


PLOTS=(ODDSRATIO(LOGBASE=E))



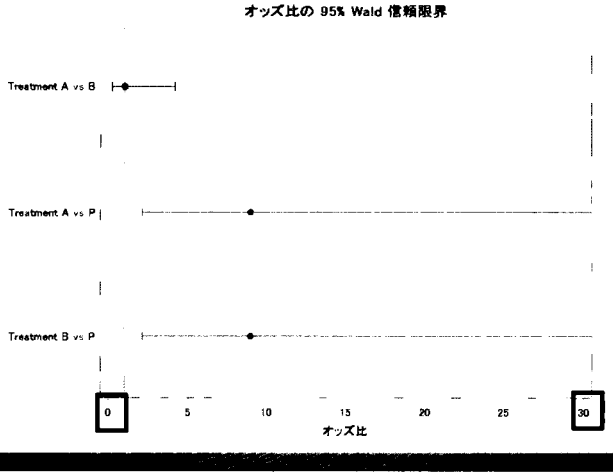
自然対数(等比スケール)，Wald型では左右対称

PLOTS=(ODDSRATIO(RANGE=CLIP))
点推定値が最小と最大の範囲



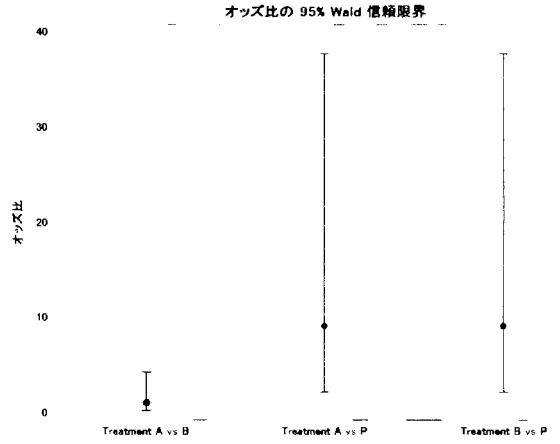
29

PLOTS=(ODDSRATIO(RANGE=(0, 30)))
オッズ比の範囲を指定



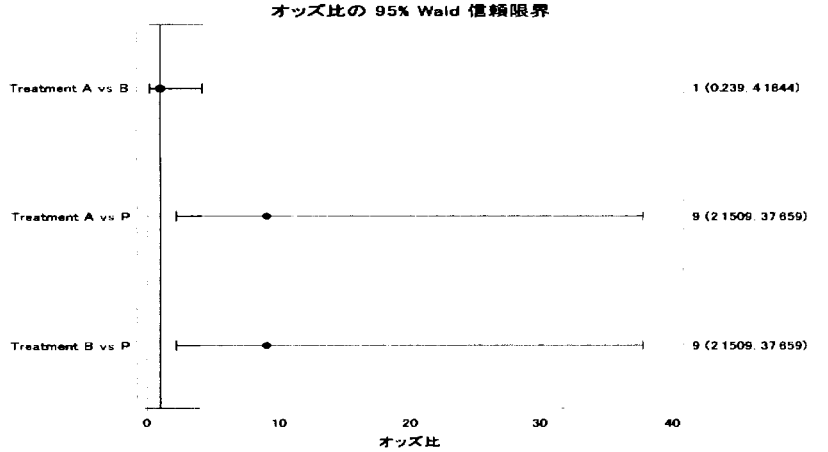
30

PLOTS=(ODDSRATIO(TYPE=VERTICAL))
垂直に信頼区間をプロット



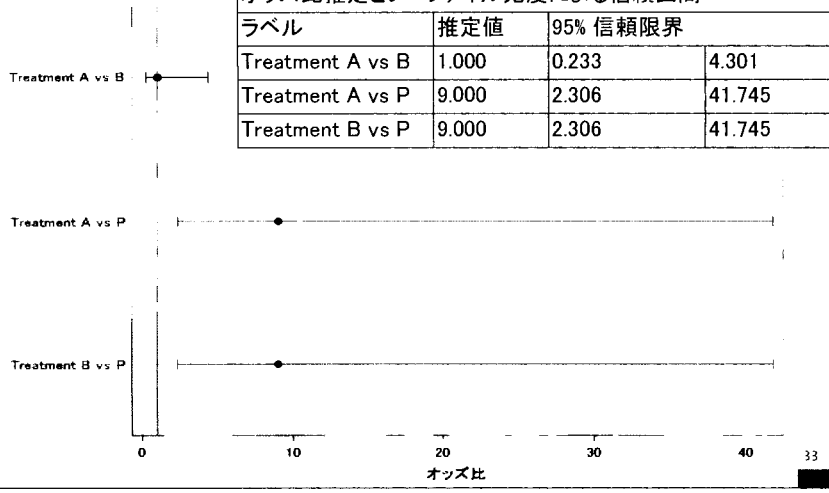
31

PLOTS=(ODDSRATIO(TYPE=HORIZONTALSTAT))
ORの数値をグラフにプロット



oddsratio treatment /ci=pl;
尤度比検定に基づく信頼区間

オッズ比推定とプロファイル尤度による信頼区間			
ラベル	推定値	95% 信頼限界	
Treatment A vs B	1.000	0.233	4.301
Treatment A vs P	9.000	2.306	41.745
Treatment B vs P	9.000	2.306	41.745



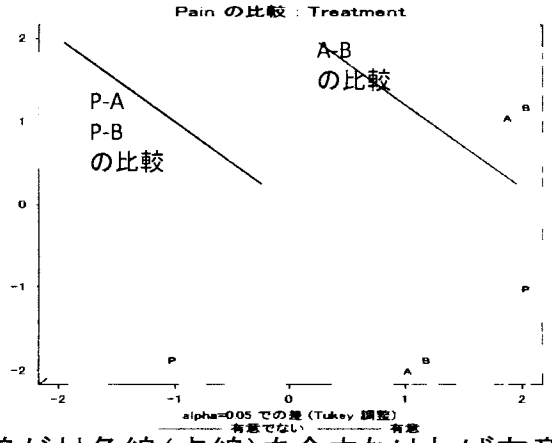
オッズ比の多重比較
lsmeans treatment/adj=tukey cl exp ;

Treatment の最小 2 乗平均の差								
多重比較の調整：Tukey					未調整オッズ比		多重性調整オッズ比	
Treatment	_Treatment	調整済 P 多重性調整p値	アルファ	指数 オッズ比	Exponentiated Lower	Exponentiated Upper	調整済 下限 Exp	調整済 上限 Exp
A	B	1	0.05	1	0.239	4.184	0.181	5.538
A	P	0.007	0.05	9	2.151	37.66	1.625	49.84
B	P	0.007	0.05	9	2.151	37.66	1.625	49.84

日本語訳の不統一はひどい

オッズ比の多重比較

lsmeans treatment/adj=tukey cl exp;



実線が対角線(点線)を含まなければ有意

オッズ比の多重比較

lsmeans treatment/adj=dunnett diff=control("P") cl exp ;

P群との比較

Treatment の最小 2 乗平均の差

多重比較の調整 : Dunnett

Treatment	_Treatment	調整済P 多重性調整p値	アルファ	指数	未調整オッズ比 Exponentiated Lower	Exponentiated Upper	多重性調整オッズ比 調整済 下限 Exp	調整済 上限 Exp
A	P	0.0051	0.05	9	2.1509	37.6593	1.7891	45.274
B	P	0.0051	0.05	9	2.1509	37.6593	1.7891	45.274

交絡を調整したオッズ比

```
ods graphics on;
proc logistic data=Neuralgia
PLOTS=(ODDSRATIO(TYPE=HORIZONTALSTAT));
   class treatment sex / param=glm;
   model pain= treatment sex age duration;
   oddsratio treatment;
run;
```


交絡を調整したオッズ比 各変数の検定結果

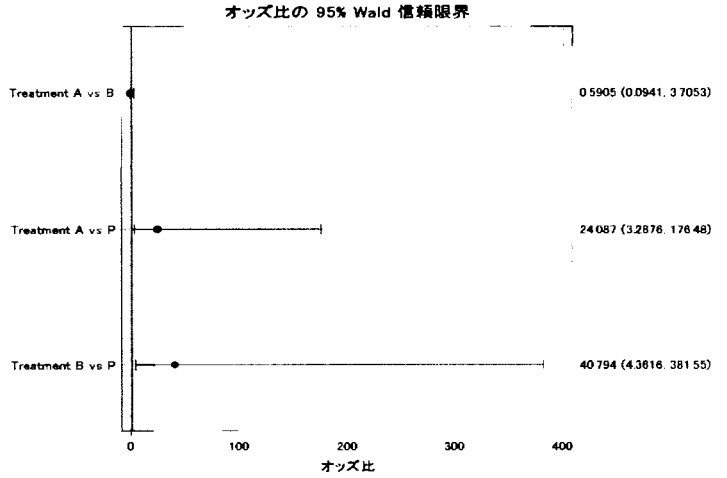
効果に対する Type 3 分析			
効果	自由度	Wald カイ 2 乗	Pr > ChiSq
Treatment	2	12.5310	0.0019**
Sex	1	5.2946	0.0214*
Age	1	7.2977	0.0069**
Duration	1	0.0315	0.8591

交絡を調整したオッズ比 パラメータ推定値

最尤推定値の分析

パラメータ		自由度	推定値	標準誤差	Wald カイ 2 乗	Pr > ChiSq
Intercept		1	15.5744	6.5915	5.5828	0.0181
Treatment	A	1	3.1817	1.0161	9.8049	0.0017**
Treatment	B	1	3.7085	1.1407	10.5700	0.0011**
Treatment	P	0	0	.	.	.
Sex	F	1	1.8322	0.7963	5.2946	0.0214*
Sex	M	0	0	.	.	.
Age		1	−0.2621	0.0970	7.2977	0.0069**
Duration		1	0.00586	0.0330	0.0315	0.8591

交絡のみを調整したオッズ比



40

多重性と交絡を調整したオッズ比

```
ods graphics on;
proc logistic data=Neuralgia
PLOTS=(ODDSRATIO(TYPE=HORIZONTALSTAT));
   class treatment sex / param=glm;
   model pain= treatment sex age duration;
   lsmeans treatment/adj=tukey cl exp;
run;
```


多重性と交絡を調整したオッズ比

Treatment の最小 2 乗平均の差 多重比較の調整：Tukey-Kramer(調整後)

		多重性調整p値					多重性調整オッズ比	
Treatment	_Treatment	Pr > |z|	調整済P	指数	Exponentiated Lower	Exponentiated Upper	調整済下限Exp	調整済上限Exp
A	B	0.5740	0.8402	0.5905	0.09409	3.7053	0.06567	5.3088
A	P	0.0017	0.0050	24.0874	3.2876	176.48	2.2261	260.64
B	P	0.0011	0.0033	40.7942	4.3616	381.55	2.8154	591.09

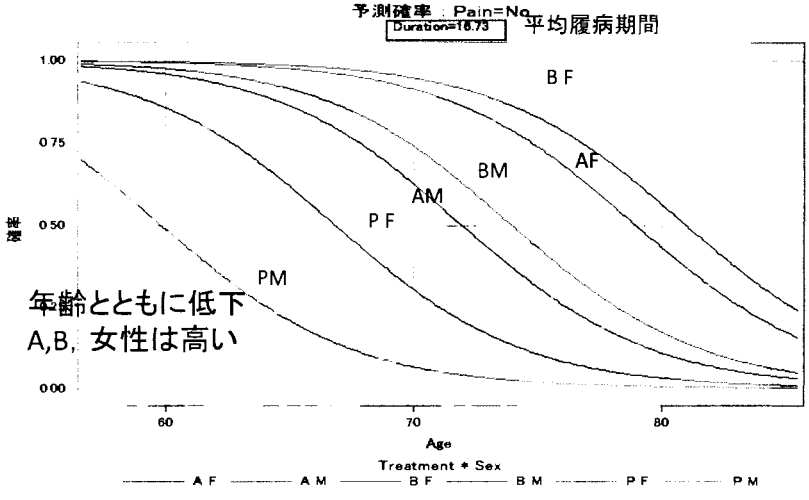
オッズ比推定と Wald による信頼区間(参考 未調整)

ラベル	推定値	95% 信頼限界	
Treatment A vs B	1.000	0.239	4.184
Treatment A vs P	9.000	2.151	37.659
Treatment B vs P	9.000	2.151	37.659

予測確率のプロット

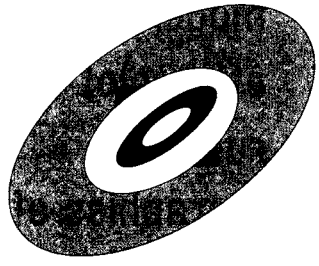
```
ods graphics on;
proc logistic data=Neuralgia
PLOTS=(EFFECT);
    class treatment sex / param=glm;
    model pain= treatment sex age duration;
    oddsratio Treatment;
run;
```


消失率の予測確率のプロット



等高線プロット

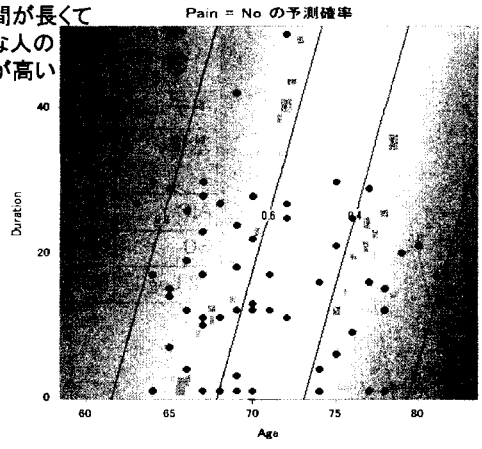
```
ods graphics on;
proc logistic data=Neuralgia;
   model pain=  age duration;
   effectplot contour;
   run;
ods graphics off;
```



等高線プロット 平面

model pain= age duration;

罹病期間が長くて
低年齢な人の
消失率が高い



●:No 消失
●:Yes 痛み有

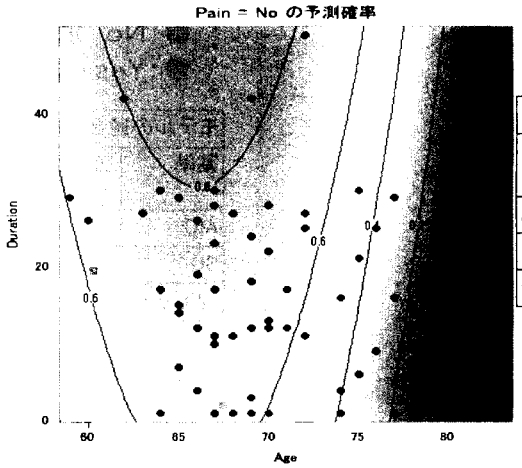
モデルの適合度統計量		
基準	切片のみ	切片と共変量
AIC	83.503	78.472
SC	85.598	84.755
-2 Log L	81.503	72.472

予測確率を
等高線で結ぶ

等高線プロット
年齢の2次の項追加

```
ods graphics on;
proc logistic data=Neuralgia;
   model pain=  age age*age duration;
   effectplot contour;
   run;
ods graphics off;
```


等高線プロット 曲面
model pain= age age*age duration;



●:No 消失
●:Yes 痛み有

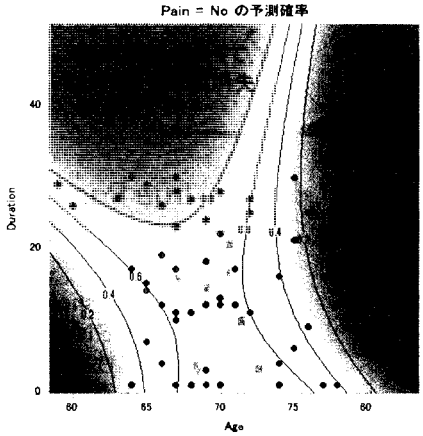
モデルの適合度統計量		
基準	切片のみ	切片と共変量
AIC	83.503	77.790
SC	85.598	86.167
-2 Log L	81.503	69.790

等高線プロット
交互作用項の追加

```
ods graphics on;
proc logistic data=Neuralgia;
   model pain= age age*age duration
duration*duration age*duration;
   effectplot contour;
   run;
ods graphics off;
```


等高線プロット 曲面

model pain= age age*age duration
duration*duration age*duration;



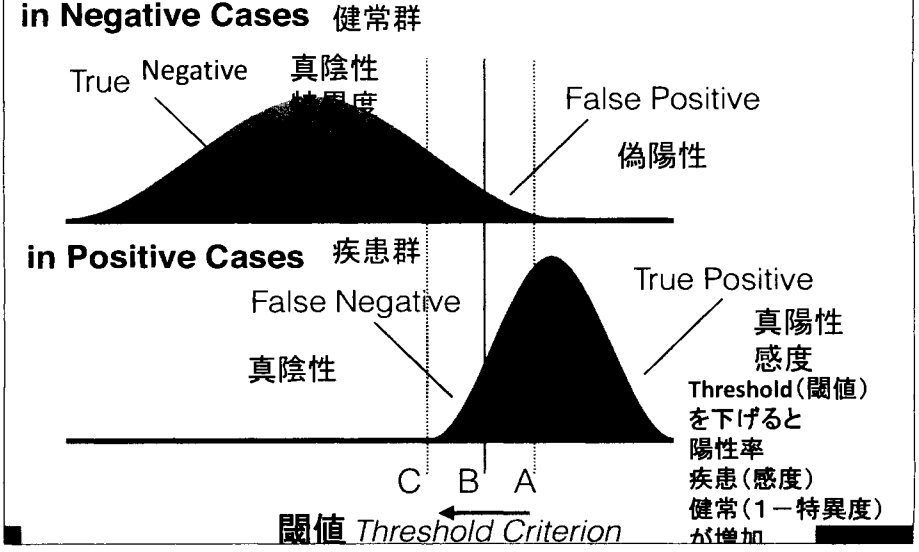
●:No 消失
●:Yes 痛み有

モデルの適合度統計量		
基準	切片のみ	切片と共変量
AIC	83.503	76.235
SC	85.598	88.801
-2 Log L	81.503	64.235

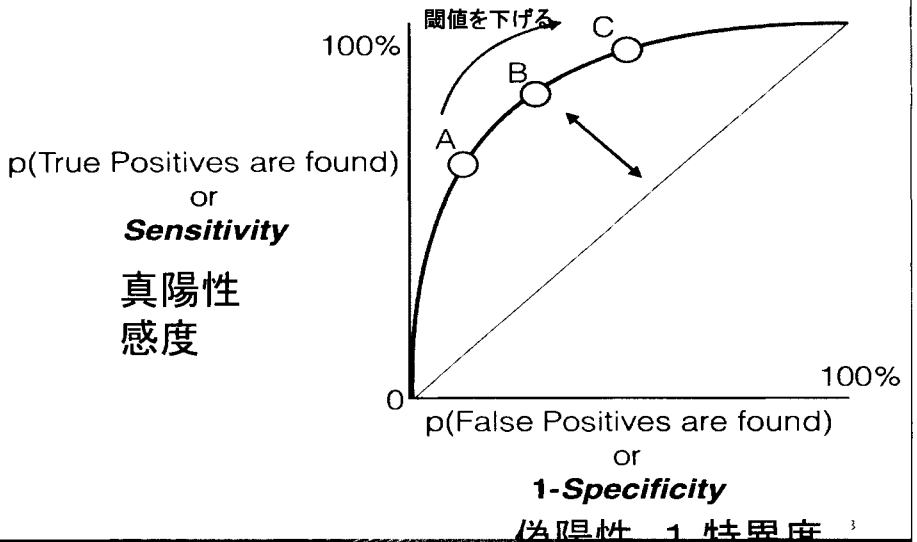
モデルの比較

モデル(切片を含むパラメータ数)	-2 Log L	AIC
duration(2)	79.886	83.886
age(2)	73.056	77.056
age duration(3)	72.472	76.472 ←
duration duration*duration age(4)	70.996	78.996
duration age age*age(4)	69.790	77.790
duration duration*duration age age*age(5)	68.360	78.360
duration duration*duration age age*age age*duration(6)	64.235	76.235 ← AIC最小

ROC(Receiver Operating Characteristic)曲線



ROC(Receiver Operating Characteristic)曲線



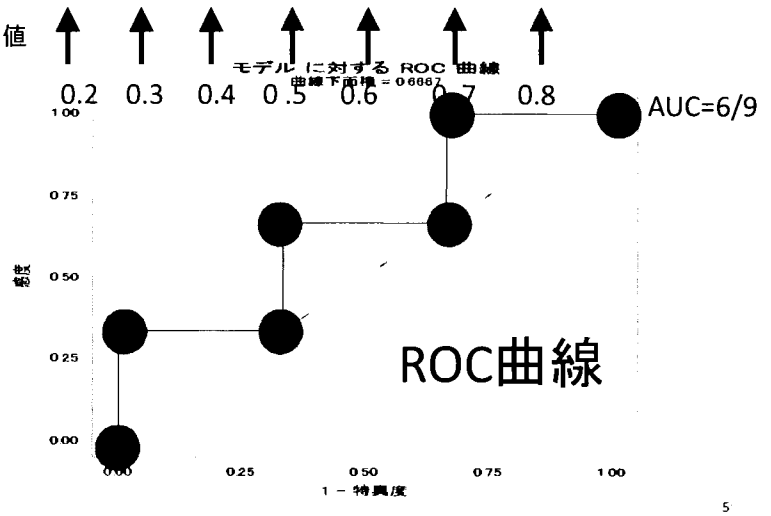
疾患なし * * *
疾患あり * * *

```
data work;
input group time censor @@;
cards;
1 4 1  1 6 1  1 8 1
2 5 1  2 7 1  2 9 1
;
proc logistic PLOTS=ROC descending;
model group=time;run;
```

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

疾患なし ＊ ＊ ＊ 1－特異度：疾患なしで閾値以上
疾患あり ＊ ＊ ＊ 感度：疾患ありで閾値以上

閾値



5

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

AUCが1に近い方が予測力が高い

（0,1）感度=1,特異度=1

感度

Mann-WhitneyのAUC？

曲線下面積(*AUC*: Area Under the Curve)
0.5～1

AUC 0.7 - 0.9	Moderate accuracy
AUC 0.5 - 0.7	Low accuracy

1－特異度

56

ROC曲線のAUCの解釈

「疾患あり」集団が「疾患なし」集団より予測確率が高くなる割合

「疾患あり」の集団から1人，「疾患なし」の集団から1人，それぞれランダムに選んだとき，「疾患あり」の人の予測確率が「疾患なし」の人の検査値よりも大きくなる確率

7

Mann-WhitneyのAUC

N_1：疾患あり群のサンプル サイズ
N_2：疾患なし群のサンプル サイズ
p_{1i}：疾患あり群の個体 iの予測確率
p_{2j}：疾患なし群の個体 jの予測確率

$N_1 \times N_2$人の予測確率の比較

$$\mathrm{AUC} = \frac{1}{N_1 \times N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} U_{ij}(p_{1i}, p_{2j})$$

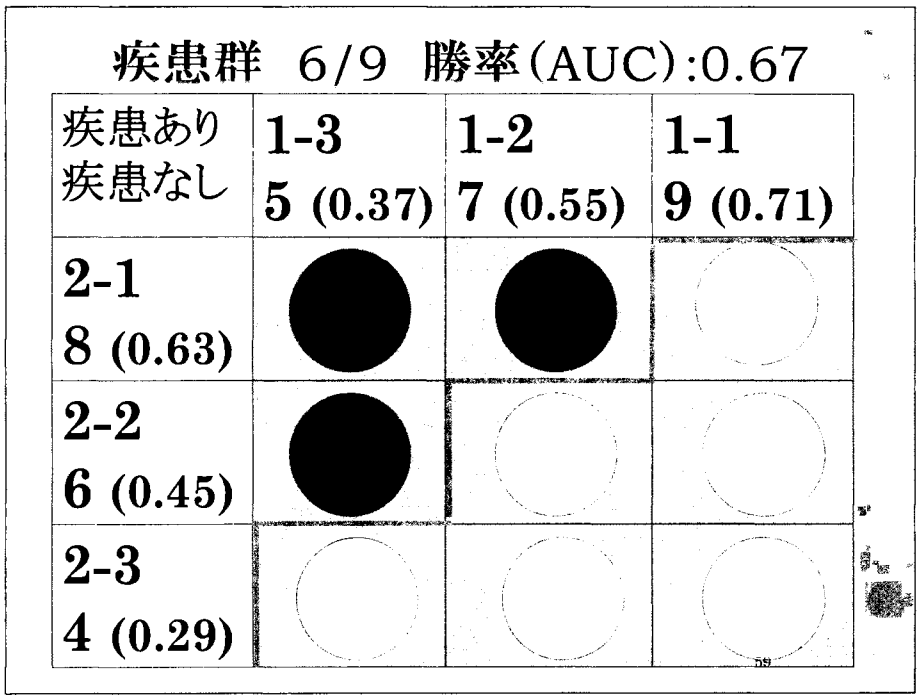
$$U_{ij}(p_{1i}, p_{2j}) = \begin{cases} 1 & p_{1i} > p_{2j} \\ .5 & p_{1i} = p_{2j} \\ 0 & p_{1i} > p_{2j} \end{cases}$$

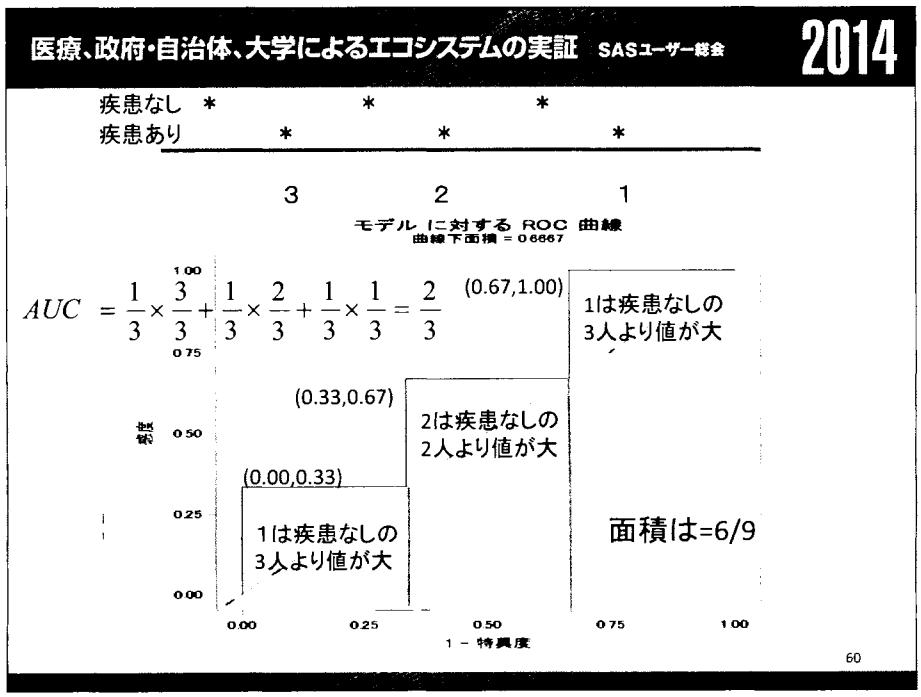
N_2人

N_1人

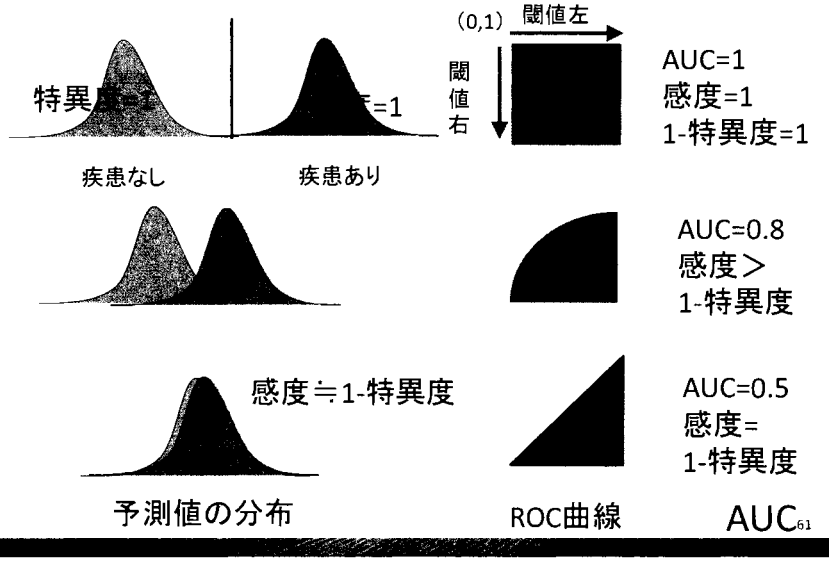
	p_{11}	p_{12}	・・・
p_{21}	0	1	
p_{22}	1	1	
・・・			

58


疾患群 6/9 勝率(AUC):0.67
疾患あり
疾患なし
1-3
5 (0.37)
1-2
7 (0.55)
1-1
9 (0.71)
2-1
8 (0.63)
2-2
6 (0.45)
2-3
4 (0.29)


医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会
2014
疾患なし
疾患あり
3
2
1
モデル に対する ROC 曲線
曲線下面積 = 0.6667
AUC = 1/3 × 3/3 + 1/3 × 2/3 + 1/3 × 1/3 = 2/3
(0.67,1.00)
1は疾患なしの
3人より値が大
(0.33,0.67)
2は疾患なしの
2人より値が大
(0.00,0.33)
1は疾患なしの
3人より値が大
面積は=6/9
感度
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 - 特異度
60

予測値の分布とAUC

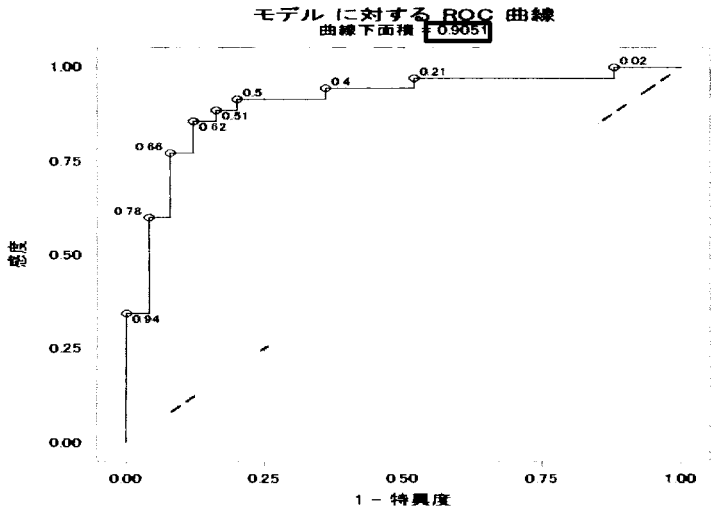


ROC曲線の作成とAUCの比較

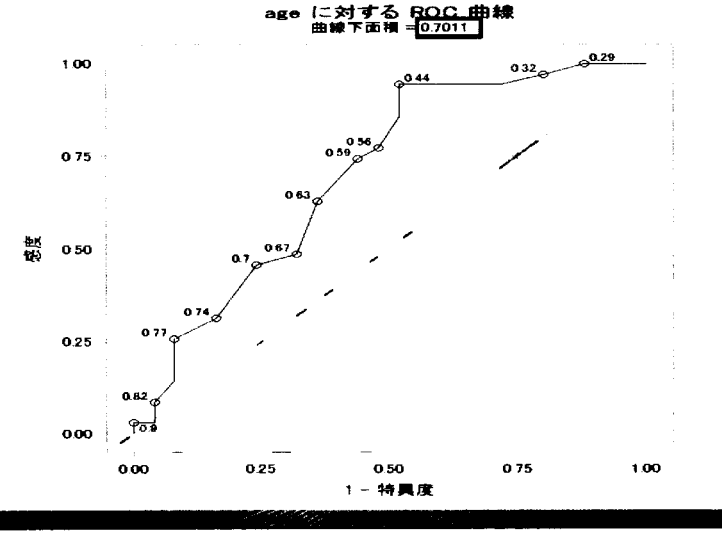
モデル(4変数)

```
ods graphics on;
proc logistic data=Neuralgia
PLOTS=ROC(ID=PROB);
   class treatment sex/param=glm;
   model pain=treatment sex age duration
   roc 'age' age ;
   roc 'duration' duration;
   roccontrast reference(model)/estimate e;
run;
```

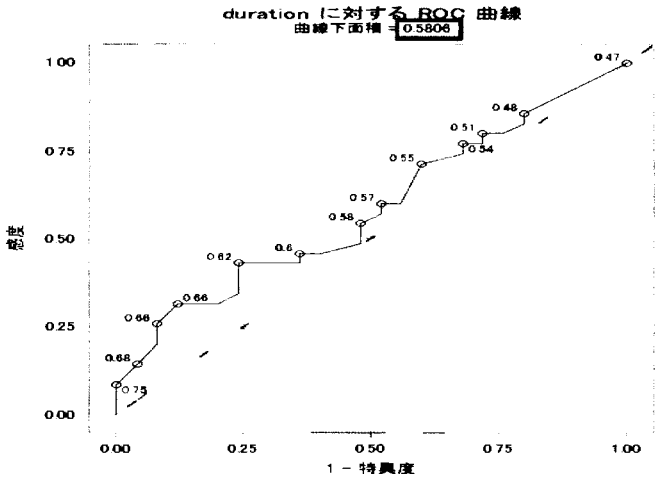
ROC曲線の作成：4変数（モデル）



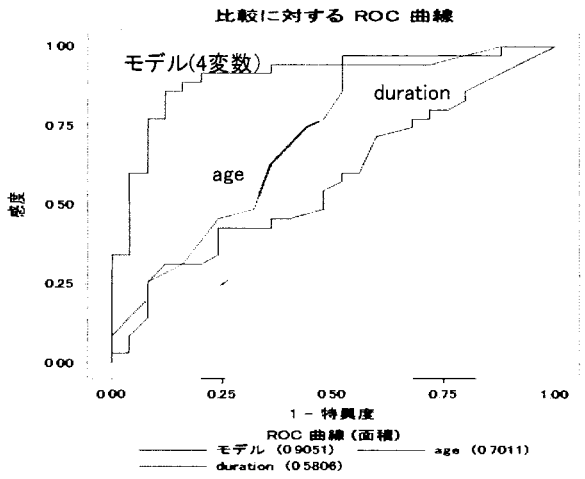
ROC曲線の作成：ageのみ



ROC曲線の作成：durationのみ



ROC曲線の比較



AUCの比較

ROC 関連性の統計量							
ROC モデル	Mann-Whitney				Somers の D (Gini)	ガンマ	Tau-a
	面積	標準誤差	95% Wald 信頼限界				
モデル (4変数)	0.9051	0.0412	0.8244	0.9858	0.8103	0.8103	0.4006
age	0.7011	0.0717	0.5605	0.8418	0.4023	0.4211	0.1989
duration	0.5806	0.0744	0.4348	0.7263	0.1611	0.1693	0.0797

AUCの比較

ROC 対比の係数		
ROC モデル	Row1	Row2
モデル	-1	-1
age	1	0
duration	0	1

ROC 対比検定の結果 3種類のモデル全体でAUCに差があるか			
対比	自由度	カイ 2 乗	Pr > ChiSq
Reference = モデル	2	20.0075	<.0001

行ごとの ROC 対比推定と検定の結果						
対比	推定値	標準誤差	95% Wald 信頼限界		カイ 2 乗	Pr > ChiSq
age - モデル	-0.2040	0.0743	-0.3496	-0.0584	7.5402	0.0060**
duration - モデル	-0.3246	0.0815	-0.4842	-0.1649	15.8789	<.0001**

0セルがある場合の例

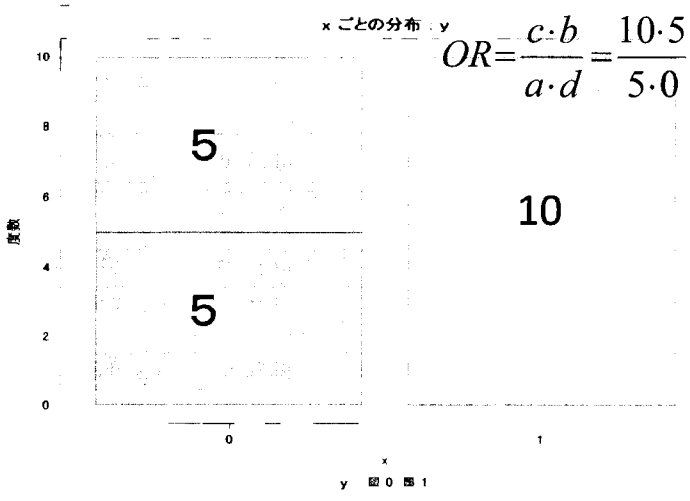
$$OR=\frac{c\cdot b}{a\cdot d}=\frac{10\cdot 5}{5\cdot 0}=\infty$$

	薬剤 −	薬剤 +	計
イベント +	5	10	15
イベント −	5	0	5
計	10	10	20

	薬剤 −	薬剤 +	計
イベント +	a	c	$n_{-\cdot}$
イベント −	b	d	$n_{+\cdot}$
計	$n_{\cdot -}$	$n_{\cdot +}$	n

c＝0のときOR＝0

オッズ比は∞，信頼区間は構成不能



$$OR=\frac{c\cdot b}{a\cdot d}=\frac{10\cdot 5}{5\cdot 0}=\infty$$

0セルがある場合の解析(FREQ)

```
data data;
input x y w @@;
cards;
0 0 5  0 1 5  1 0 0  1 1 10
;
ods graphics on;
proc freq;
tables x*y/
plots=freqplot(TWOWAY=stack)cmh;
weight w;run;
```


CMHオプション

$$OR=\frac{5.5\cdot 10.5}{5.5\cdot 0.5}=\frac{10.5}{0.5}=21$$

相対リスクの推定値（行 1 / 行 2）				
研究の種類	調整方法	値	95% 信頼限界	
ケースコントロール研究	Mantel-Haenszel	.	.	.
(オッズ比)	ロジット **	21.0000	0.9716	453.9116
コーホート研究	Mantel-Haenszel	.	.	.
(列 1 のリスク)	ロジット **	11.0000	0.6880	175.8626
コーホート研究	Mantel-Haenszel	0.5000	0.2690	0.9293
(列 2 のリスク)	ロジット	0.5000	0.2690	0.9293

計算していない推定値があります。
** セル度数が 0 を含む表の場合には、その表の全セルに0.5を加えて、調整を行っています。

セル度数＋0．5

$$OR=\frac{c\cdot b}{a\cdot d}=\frac{5.5\cdot 10.5}{5.5\cdot 0.5}=21.0$$

	薬剤－	薬剤＋	計
イベント＋	5.5	10.5	11
イベント－	5.5	0.5	6
計	11	11	202

	薬剤－	薬剤＋	計
イベント＋	a+.5	c+.5	$n_{-.}$+1
イベント－	b+.5	d+.5	$n_{+.}$+1
計	$n_{.-}$+1	$n_{.+}$+1	n+2

LOGISTICプロシジャによる解析

```
proc logistic descending;
model y=x ;freq w;run;
proc logistic descending;
model y=x /firth;freq w;run;
proc logistic descending;
model y=x /;exact x/estimate=both;
freq w;run;
```

$$p=\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$$

通常の解析 $p = \frac{\exp(\infty \cdot x)}{1 + \exp(\infty \cdot x)}$

最尤推定値の分析					
パラメータ	自由度	推定値	標準誤差	Wald カイ 2 乗	Pr > ChiSq
Intercept	1	-106E-18	0.6325	0.0000	1.0000
x	1	12.5661	169.3	0.0055	0.9408

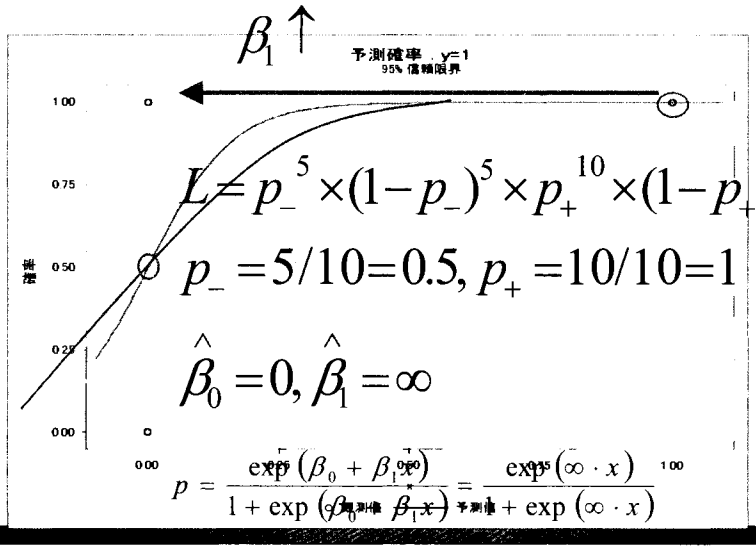
オッズ比の推定			
効果	点推定値	95% Wald 信頼限界	
x	>999.999	<0.001	>999.999

WARNING: 恐らくデータ点が準完全分離の状態です。最尤推定値は存在しないかもしれません。
WARNING: LOGISTICプロシジャは上記の警告にもかかわらず継続します。
最尤反復に基づいて結果が表示されます。モデルの当てはめの妥当性は疑わしいです。

$\beta_0 \to 0 \quad \beta_1 \to +\infty$

最大尤度反復履歴				
反復	リッジ	-2 対数尤度	Intercept $\hat{\beta}_0$	x $\hat{\beta}_1$
0	0	22.493406	1.098612	0
1	0	15.684718	-0.234721	2.666667
2	0	14.446460	0.002161	3.517650
3	0	14.073301	-1.682489E-9	4.549416
4	0	13.939772	6.938894E-17	5.559990
5	0	13.891133	-1.07553E-16	6.563839
6	0	13.873304	7.112366E-17	7.565249
7	0	13.866754	-1.06252E-16	8.565767
8	0	13.864345	7.123208E-17	9.565958
9	0	13.863459	-1.0636E-16	10.566028
10	0	13.863133	7.127274E-17	11.566054
11	0	13.863013	-1.06353E-16	12.566063

準完全分離:complete separation

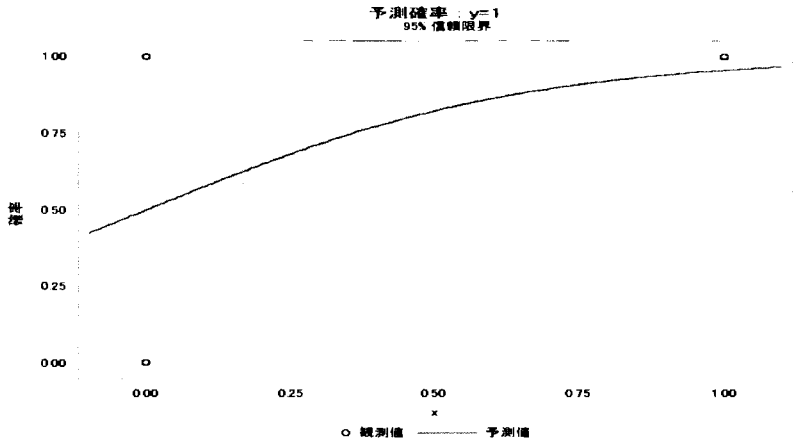


Firth法:パラメータ推定値

最尤推定値の分析					
パラメータ	自由度	推定値	標準誤差	Wald カイ 2 乗	Pr > ChiSq
Intercept	1	-0.00003	0.6325	0.0000	1.0000
x	1	3.0446	1.6446	3.4270	0.0641

オッズ比の推定			
効果	点推定値	95% Wald 信頼限界	
x	21.001	0.836	527.438

Firth法：予測確率と信頼区間



ロジスティック回帰

	−	+
イベント +	5	10
イベント −	5	0

likelihood（尤度）

薬剤ー:x=0,薬剤+:x=1

尤度（L）＝モデルの下でデータが得られる確率

$$L = p_-^{5} \times (1-p_-)^{5} \times p_+^{10} \times (1-p_+)^{0}$$

$$p_- = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}, \quad p_+ = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}$$

最尤法：β_0, β_1の値を動かしてLが最も大きくなるようにする方法

MLE : Maximum Likelihood Estimator

イベント+	5	10
イベント−	5	0

対数尤度とスコア関数

$$L = p_-^{5} \times (1-p_-)^{5} \times p_+^{10} \times (1-p_+)^{0}$$

$$\log L = 5\log p_- + 5\log(1-p_-) + 10\log p_+ + 0\log(1-p_+)$$

$$U(\beta_1) = \frac{d\log L}{d\beta_1} = 10-(10+0)p_+ = 0 \Rightarrow p_+ = \frac{10}{10+0}$$

$$U(\beta_0) = \frac{d\log L}{d\beta_0} = 5+10-(5+5)p_- -(10+0)p_+ = 0 \Rightarrow p_- = \frac{5}{5+5}$$

イベント+	5	10
イベント−	5	0

最尤推定量

$$\log\frac{p}{1-p} = \beta_0 + \beta_1 x$$

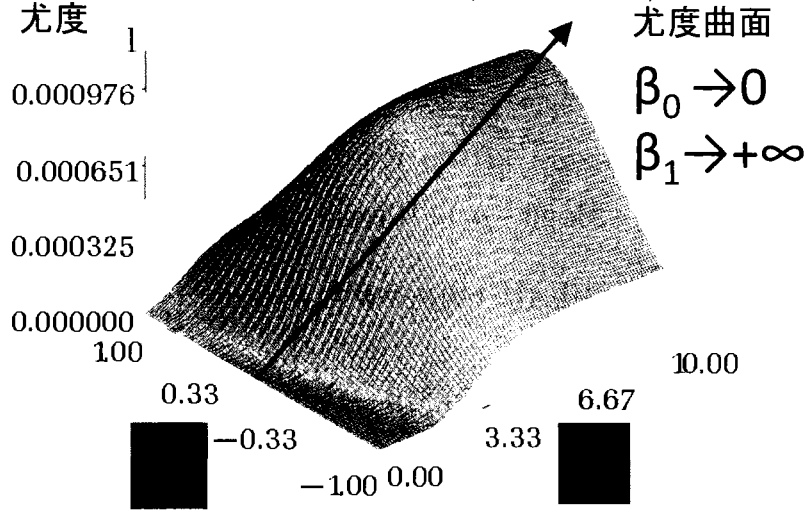
$$\hat{\beta}_0 = \log\frac{p_-}{1-p_-} = \log\frac{5/(5+5)}{1-5/(5+5)} = \log\frac{5}{5} = 0$$

$$\hat{\beta}_0 + \hat{\beta}_1 = \log\frac{p_+}{1-p_+} = \log\frac{10/(10+0)}{1-10/(10+0)} = \log\frac{10}{0} = \infty$$

$$\hat{\beta}_1 = \log\frac{p_+}{1-p_+} - \log\frac{p_-}{1-p_-} = \log\left(\frac{10\cdot 5}{5\cdot 0}\right) = \infty$$

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

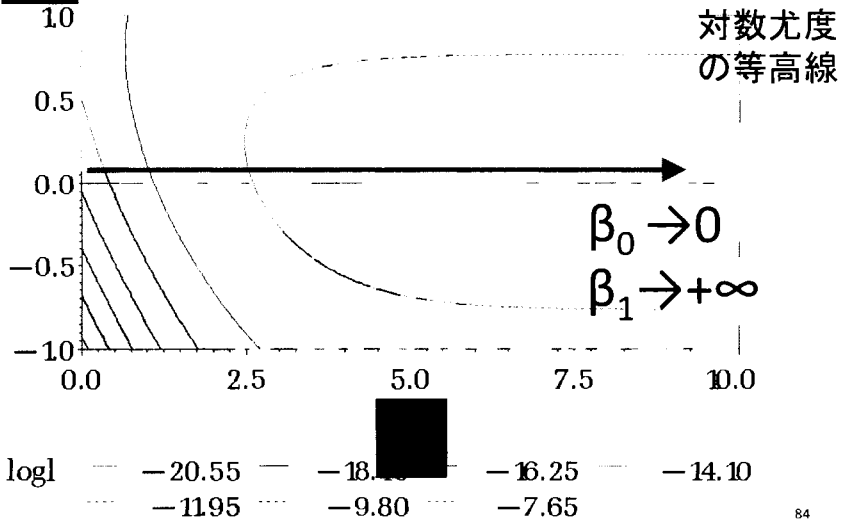
$$L = p_-^{5} \times (1-p_-)^{5} \times p_+^{10} \times (1-p_+)^{0}$$



83

医療、政府・自治体、大学によるエコシステムの実証 SASユーザー総会 2014

$$\log L = \log\left(p_-^{5} \times (1-p_-)^{5} \times p_+^{10} \times (1-p_+)^{0}\right)$$



84

Firth法：一般論
尤度に罰則項を積算

尤度： $\mathrm{L}(\boldsymbol{\beta} \mid \mathbf{Y}), U(\boldsymbol{\beta})$：スコア関数 ,情報行列： $I(\boldsymbol{\beta})$

罰則項付き尤度： $\mathrm{L}^*(\boldsymbol{\beta} \mid \mathbf{Y}) = \mathrm{L}(\boldsymbol{\beta} \mid \mathbf{Y}) \times \boxed{|I(\boldsymbol{\beta})|^{0.5}}$

$\log \mathrm{L}^*(\boldsymbol{\beta} \mid \mathbf{Y}) = \log \mathrm{L}(\boldsymbol{\beta} \mid \mathbf{Y}) + 0.5 \log |I(\boldsymbol{\beta})|$

$|I(\boldsymbol{\beta})|^{0.5}$: 罰則項 , 情報行列の行列式の 0.5乗

$$U^*(\boldsymbol{\beta}) = U(\boldsymbol{\beta}) + 0.5 tr\left[I(\boldsymbol{\beta})^{-1} \frac{dI(\boldsymbol{\beta})}{d\boldsymbol{\beta}} \right]$$

単変量のモデル： $p_i = \dfrac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$

85

$$L = \left(\frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right)^{y_i} \left(\frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \right)^{1 - y_i}$$

$$\log L = \sum y_i (\beta_0 + \beta_1 x_i) - \sum \log\left[1 + \exp(\beta_0 + \beta_1 x_i)\right]$$

$$U_0 = \frac{d \log L}{d\beta_0} = \sum y_i - \sum \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

$$= \sum (y_i - p_i)$$

$$U_1 = \frac{d \log L}{d\beta_1} = \sum x_i y_i - \sum \frac{x_i \exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

$$= \sum x_i (y_i - p_i)$$

$$\mathbf{U}^T = \sum (y_i - p_i)\mathbf{x}_i = \left[\sum (y_i - p_i) \quad \sum x_i (y_i - p_i) \right]$$

情報行列

$$\mathbf{I}=-\frac{d\log L}{d\boldsymbol{\beta}d\boldsymbol{\beta}^{T}}$$

$$\begin{bmatrix} -\frac{d^2\log L}{d\beta_0 d\beta_0} & -\frac{d^2\log L}{d\beta_0 d\beta_1} \\ -\frac{d^2\log L}{d\beta_1 d\beta_0} & -\frac{d^2\log L}{d\beta_1 d\beta_1} \end{bmatrix} = \begin{bmatrix} \sum p_i(1-p_i) & \sum x_i p_i(1-p_i) \\ \sum x_i p_i(1-p_i) & \sum x_i^2 p_i(1-p_i) \end{bmatrix}$$

推定値の分散は情報行列
の逆行列によって与えられる.

情報行列(2×2の分割表) x_iは0,1

$$\mathbf{I}=\begin{bmatrix} \sum p_i(1-p_i) & \sum x_i p_i(1-p_i) \\ \sum x_i p_i(1-p_i) & \sum x_i^2 p_i(1-p_i) \end{bmatrix}=\boxed{\begin{bmatrix} e & f \\ f & f \end{bmatrix}}$$

$$e=(a+b)p_-(1-p_-)+(c+d)p_+(1-p_+)$$

$$f=(c+d)p_+(1-p_+)$$

$$|I|=ef-f^2=f(e-f)=(c+d)p_+(1-p_+)(a+b)p_-(1-p_-)$$

$$=\boxed{n_+n_-p_+(1-p_+)\cdot p_-(1-p_-)}$$

Firth法のpenalized尤度L*
尤度と情報量の積を最大化

$$L^* = L \times |I|^{1/2}$$

$$L = p_-^{\,a} \times (1-p_-)^b \times p_+^{\,c} \times (1-p_+)^d$$

$$|I| = (c+d)p_+(1-p_+)(a+b)p_-(1-p_-)$$

$$L^* \propto p_-^{\,a} \times (1-p_-)^b \times p_+^{\,c} \times (1-p_+)^d \times [p_+(1-p_+)p_-(1-p_-)]^{1/2}$$

$$= p_-^{\,a+.5} \times (1-p_-)^{b+.5} \times p_+^{\,c+.5} \times (1-p_+)^{d+.5}$$

$$OR = \exp(\hat{\beta}_1) = \frac{(b+0.5)(c+0.5)}{(a+0.5)(d+0.5)}$$

Firth法のpenalized尤度L*
0セルが存在する場合(d=0)

$$L^* = L \times |I|^{1/2} = L \times V[y_+] \times V[y_-]$$

$$L = p_-^{\,a} \times (1-p_-)^b \times p_+^{\,c} \times (1-p_+)^d$$

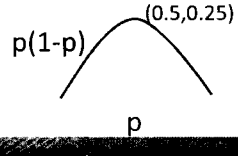
$$= p_-^{\,a} \times (1-p_-)^b \times p_+^{\,c}$$

$$|I| = (c+d)p_+(1-p_+)(a+b)p_-(1-p_-)$$

$$= cp_+(1-p_+)(a+b)p_-(1-p_-)$$

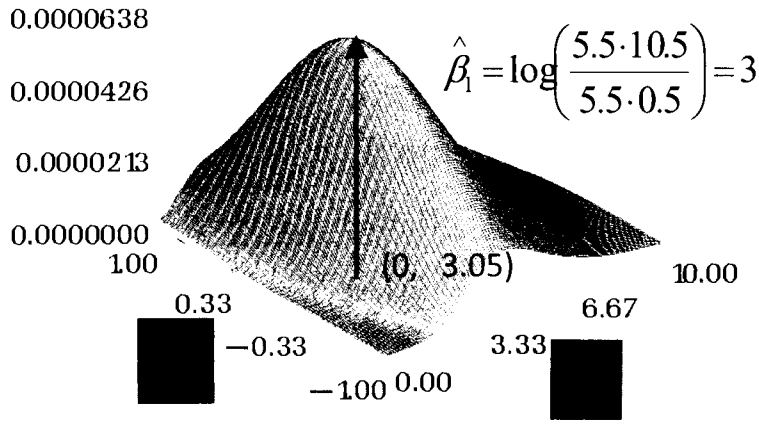
$$Max \quad L \Rightarrow p_+ = 1$$

$$Max \quad |I| \Rightarrow p_+ = 0.5$$



$$L = p_{-}^{5.5} \times (1-p_{-})^{5.5} \times p_{+}^{10.5} \times (1-p_{+})^{0.5}$$

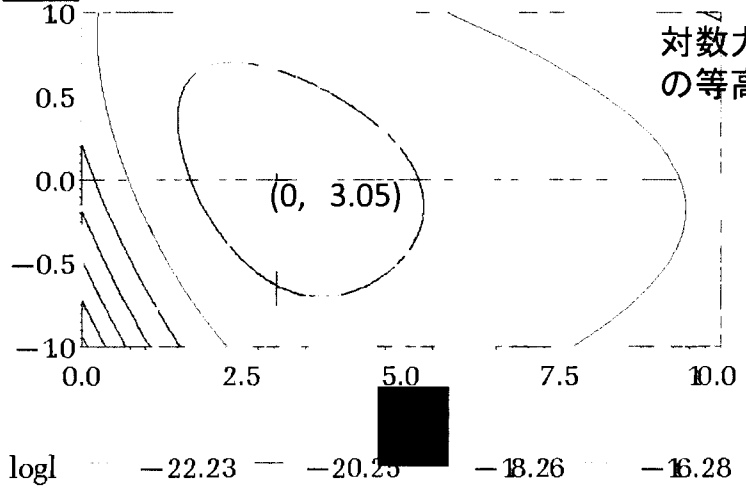
尤度曲面



$$\hat{\beta}_1 = \log\left(\frac{5.5 \cdot 10.5}{5.5 \cdot 0.5}\right) = 3.05$$

$$\log L = \log\left(p_{-}^{5.5} \times (1-p_{-})^{5.5} \times p_{+}^{10.5} \times (1-p_{+})^{0.5}\right)$$

対数尤度の等高線



田中勇輔　浜田知久馬　佐野雅隆(2014)
稀なイベントを対象としたメタ・アナリシスの性能評価　計量生物学会

0セル存在下のメタアナリシス

通常の方法でオッズ比を算出すると，

2群の一方で0セルがあるとメタアナリシスの対象外，オッズ比は0または無限大

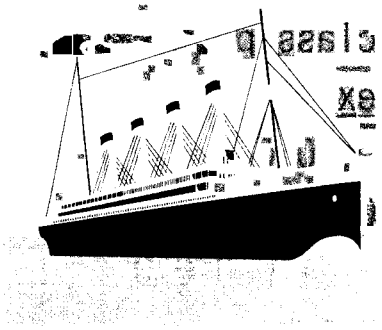
安全性のメタアナリシス

ロシグリタゾン（心筋梗塞）では(30/42)研究が除外

ベバシズマブ（治療関連死）では(2/14)研究が除外

93

JackとRoseの
ロマンティック回帰



「運命を変える恋がある．」

しかし2人の死亡確率は

大きく違った．

ジャックは20歳男性で
3等船室の乗客

ローズは20歳女性で
1等船室の乗客

http://www.encyclopedia-titanica.org/index.php
タイタニック号の乗員のデータベース

94

2人+ローズ婚約者の死亡確率

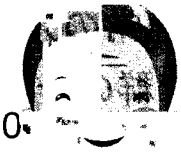
ジャックは20歳男性で3等船室の乗客

$$\log\left(\frac{p}{1-p}\right) = -3.5221 + 2.4978 + 2.2897 + 0.0$$

p=0.95

ローズは20歳女性で1等船室の乗客

$$\log\left(\frac{p}{1-p}\right) = -3.5221 + 0.0344 \times 20$$



p=0.20

ローズ婚約者は男性で1等船室の乗客

$$\log\left(\frac{p}{1-p}\right) = -3.5221 + 2.4978 + 0.0344 \times 20$$

p=0.45

$$\log\left(\frac{p}{1-p}\right) = -3.5221 + 2.4978SEX(MALE) + 1.2806PCLASS(2) + 2.2897PCLASS(3) + 0.0344AGE$$

男＞女　2等＞1等　3等＞1等

Jack, Rose, Rose婚約者の死亡のオッズ比

```
proc logistic data=titanic ;
class sex pclass/param=glm ref=first;
model survived=sex pclass pclass*sex;
lsmestimate pclass*sex
   " Jack-Rose"     -1 0 0 0 0 1,
   " Jack-fiance"   0 -1 0 0 0 1,
   " fiance-Rose"   -1 1 0 0 0 0
  /adj=simulate exp;run;
```

R F J

シミュレーションによって，任意の対比の多重性の調整が可能

1	1	2	2	3	3
女	男	女	男	女	男

Jack, Rose, Rose婚約者の死亡のオッズ比の多重比較

Bonferroni法ほど保守的ではない.

Least Squares Means Estimates
Adjustment for Multiplicity: Simulated 多重性調整p値

効果	ラベル	推定値	標準誤差	z 値	Pr > |z|	調整済P	指数
SEX*PCLASS	Jack-Rose	5.0430	0.4721	10.68	<.0001	<.0001	154.94
SEX*PCLASS	Jack-fiance	3.7764	0.3312	11.40	<.0001	<.0001	43.6578
SEX*PCLASS	fiance-Rose	1.2666	0.5488	2.31	0.0210	0.0516	3.5489

ラブロマンスと生存パターン

ヒーロー	ヒロイン	ラブロマンス	確率
生存	生存	シンデレラ 美女と野獣	0.05×0.80 =0.04
生存	死亡	ある愛の詩 野菊の墓 赤い疑惑	0.05×0.20 =0.01
死亡	生存	タイタニック 愛と誠 続星の金貨	0.95×0.80 =0.76
死亡	死亡	ロミオとジュリエット 失楽園	0.95×0.20 =0.19

参考文献

Christopher Zorn (2005)
A Solution to Separation in Binary Response Models. Political Analysis 13:157–170

大倉征幸・鎌倉稔成(2007)
精確ロジスティック回帰の近似推定値.
応用統計学,36-2,3, 87-98

TUTORIAL

政府統計ミクロデータの符号表からSAS変数のラベルとフォーマットを自動生成するSASプログラムの作成方法

How to Write a SAS Program to Automatically Create Labels and Formats for SAS Variables from a "Code Table" Provided together with Public Micro Data of Japanese Government

周防　節雄（兵庫県立大学・名誉教授）

要旨

日本の政府機関が作成したミクロデータには、メタデータとして「符号表」が付随している。この符号表は、政府機関共通の「政府統計個票データレイアウト標準記法」（平成 18 年）に準拠して作成されており、エクセルファイルでネットからもアクセス・保存が出来る。そこには、変数名、変数の長さ、変数の型（数値／文字）、ラベル、変数コードとその内容等が階層構造で表現されている。このエクセルファイルからSAS変数のラベルとSASフォーマットを自動生成し、かつ、SAS変数も自動的にrenameするSASプログラムについて解説する。本チュートリアルでは、SASユーザー総会のデータ分析コンペに使用されている平成 16 年全国消費実態調査の教育用擬似ミクロデータの符号表を例題として用いるので、今後データコンペに参加予定のユーザーには大いに参考になるはずである。

1. はじめに

近年、パソコンの急速な高性能化と、データ分析用の優秀なソフトウェアの整備のおかげで、以前ではメインフレームコンピュータでしか出来なかった公的統計[1]のミクロデータ分析が大学・研究機関の研究室レベルで盛んに行われるようになってきた。公的統計のミクロデータの利用は、平成 19 年に新「統計法」が制定される以前は、ミクロデータの「目的外使用」と一般的に呼称されていたが、今では「調査票情報の二次利用」[2]と言われている。

ミクロデータの提供の際は、以前は各省庁がそれぞれ独自のファイルレイアウトを使っていた。平成 18 年に政府機関共通の「政府統計個票データレイアウト標準記法」が制定され、現在では、ミクロデータ本体と共に、この標準記法に準拠した「レイアウト表」と「符号表」がエクセルファイルで提供される。符号表（付録１）に含まれる主な情報としては、項目名、長さ、型（数値型、文字型）、変数名、符号（いわゆるコード値）とその内容（意味）がある。

データ分析の過程では、変数のラベルとフォーマットが不可欠であるが、ミクロデータには変数の数が多いのが通例で、その作成には大変な手間と時間がかかる。そこで、手作業で変数のラベルとフォーマットを作成する代わりに、符号表を入力データとしてこれらをSASプログラムで自動的に作成

---

1 公的統計とは、国・地方公共団体やその他の公的機関が作成する統計を指し、以前は「政府統計」とか「官庁統計」と呼称していた。公的統計のわかりやすい解説本として、『公的統計の体型と見方』（松井博 2008）がある。

2 統計法（平成十九年五月二十三日法律第五十三号）の第三十二条以下に関連の規定がある。統計法の条文は以下の URL で閲覧出来る。http://law.e-gov.go.jp/htmldata/H19/H19H0053.html

すれば、便利だと考えた。本チュートリアルでは、SAS ユーザー総会で 2013 年からデータ分析コンペで使用されている平成 16 年全国消費実態調査の教育用擬似ミクロデータの符号表を例題として用いているが、このプログラムは他の公的統計のミクロデータの符号表にもそのまま適用できるので、利用者には役に立つプログラムと自負している。

2. 全国消費実態調査の教育用擬似ミクロデータ

2.1 概要

全国消費実態調査の教育用擬似ミクロデータ(以下「ミクロデータ」と呼称する)については「教育用擬似ミクロデータの開発とその利用 ～平成 16 年全国消費実態調査を例として～」と題する文書が以下の URL で閲覧出来るので、詳細についてはそちらを参照して欲しいが、ここでは同文書の「2.2 教育用擬似ミクロデータ」を以下に引用しておく。

http://www.nstac.go.jp/services/pdf/sankousiryou2407.pdf

> 調査票情報から作成したものは、調査票情報であることを踏まえて、教育用擬似ミクロデータでは、個票データから高次元の集計表を作成し、その高次元の集計表から個票データに近似したミクロデータを作成するという方法をとっている。集計表から作成するために、個票データでも、匿名データでもない擬似的なミクロデータと言える。それでいて、この教育用ミクロデータは、実証分析に利用した際に、我が国の実態を反映できるように、つまり個票データの分布にできる限り近似するように工夫して作成する方向で考えた。
>
> このように集計表から作成する教育用擬似ミクロデータは、基本的に、①個票データの分布に近づけるなど、元の個票データに近似したデータであること、②量的属性の相関関係を保つなど、量的属性間の関係が整合的であること、③全国消費実態調査で言えば収入総額と支出総額が合致しているなど、調査特有のデータ構造を保持すること、④標本調査における集計用乗率を考慮すること、⑤データ量は元の個票データに合わせること、の考えの下で作成している。作成例としての全国消費実態調査における考慮点として、質的属性 5 については、集計表の作成における分類項目が該当し、その項目数は限られたものになり、量的属性については、分析上必要と思われる収入項目、支出項目を収録する。

2.2 教育用擬似ミクロデータの SAS データセットへの変換

本チュートリアルの議論に必要なので、提供されるミクロデータ本体の SAS データセットへの変換方法について簡単に触れておきたい。

提供されたミクロデータは 197 変数、32,028 レコードから成る CSV ファイルであるが、提供に際しては 7 つの CSV ファイルに分割されている。最初のレコードから直ぐにデータが始まっており、いわゆる項目名とか変数名に相当するメタ情報はない。このような場合、SAS データセットへ変換するには様々な方法があるが、一番簡単にしかも確実に変換するには、proc import を使うのが最適である。

当初、7 つの CSV ファイルをそれぞれ先にエクセルファイルに変換してから、proc import のマクロを作って、7 つの SAS データセットに取り込んだ後に、その 7 個を set 文で縦に結合しようとした。ところが、proc import を使うと、文字変数の長さは、各ファイル内で最大の長さが自動的に定義されるので、データセットによっては同じ名前の変数なのに length が異なるケースが出て、結合しようと

すると、エラーになり、うまくいかなかった。どうしても、proc import を使いたかったので、まず、DOS コマンドで事前に CSV ファイル 7 個を縦に結合してから、proc import で SAS データセットに変換した。その結果、F1～F197 の変数、32,028 オブザベーションから成るデータセットが作成できた。

このデータセットをそのまま使うのでは、各変数が何を意味しているかわかりにくく、分析過程では大変な不便を感じるし、何よりも重大な変数の取り違えが起こりかねない。そこで、少なくともよく使う変数はそれと分かる変数名(mnemonic name)に変えて、かつ、変数ラベルも付けておくのが望ましい。この変数名の mnemonic 化、変数ラベルの自動設定及び、変数の format の作成をする方法を以下で解説する。

2.3 符号表と SAS データセット化

ミクロデータの提供を受けると、ミクロデータ本体の他に、メタデータとして、「レイアウト表」と「符号表」がそれぞれのシートに保存されたエクセルファイルも提供される。本チュートリアルでは、この符号表から、変数のラベル・フォーマットの作成、及び変数の mnemonic 化をする SAS プログラムを論じているので、まず、符号表の中味に限定して概要を解説する。

提供された「平成 16 年全国消費実態調査符号表」の先頭部分を付録 1 に示す。このエクセルファイルのオリジナル版にはセキュリティ用のパスワードが設定されており、そのままでは SAS プログラムで読めないので、パスワード設定を外したファイルに変換して使う必要がある。そのファイルに対しては、proc import で SAS データセットに取り込む前に加工を施している(付録 2)が、詳細は、3.2 節で述べている。このエクセルファイルをそっくり同じイメージで SAS データセットに取り込んだ結果を表示した「結果ビューア画面」を付録 3 に示す。

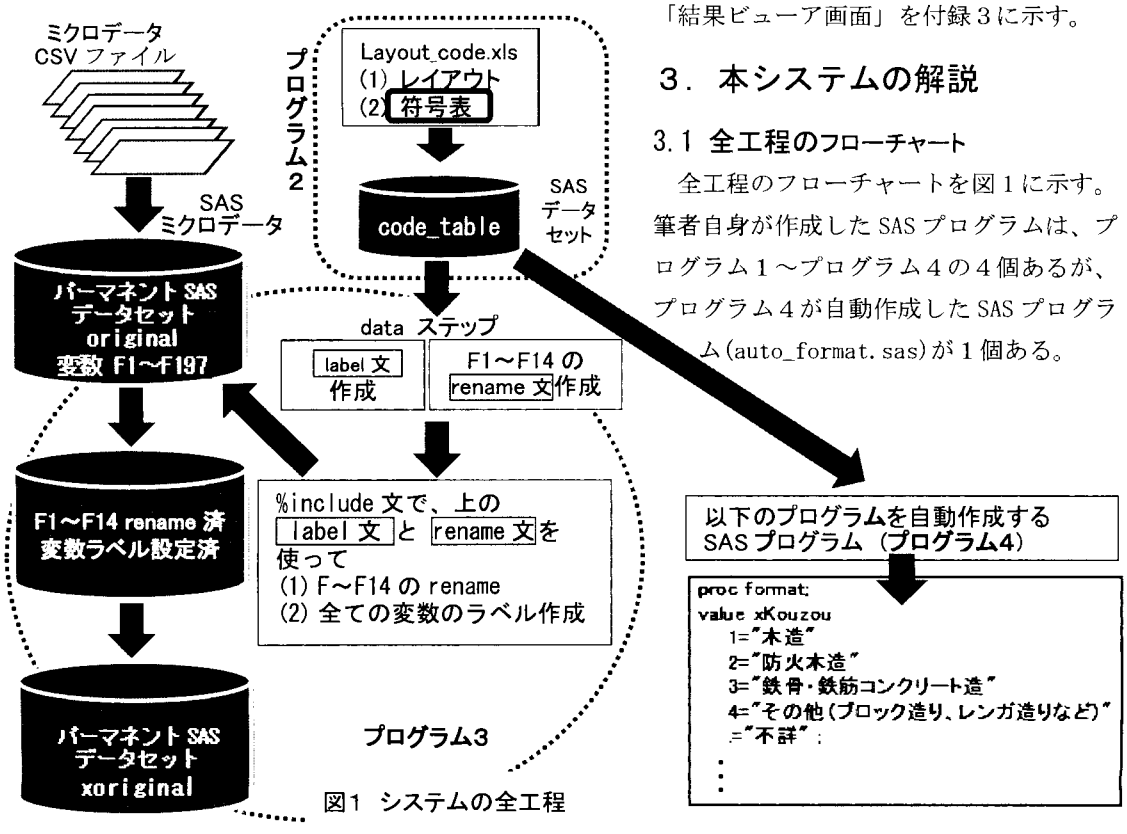


図1 システムの全工程

3. 本システムの解説

3.1 全工程のフローチャート

全工程のフローチャートを図 1 に示す。筆者自身が作成した SAS プログラムは、プログラム 1～プログラム 4 の 4 個あるが、プログラム 4 が自動作成した SAS プログラム(auto_format.sas)が 1 個ある。

3.2 符号表をSASデータセットに変換

プログラム１でプログラムの動作環境を設定する。利用者は自分のパソコン環境に応じて、ここで設定しておけば、以下に示す一連のプログラムで再度設定は不要である。

エクセル形式で統計センターから提供される符号表(付録１)の先頭部には符号表に関する情報や、各列の項目名があるが、proc importを使う際には、余分なレコードなので削除して、その代わりに、項目名に相当するSAS変数名を入力しておく(付録２)。プログラム２を実行すると、このエクセルファイルがパーマネントSASデータセットcode_table(付録３)に変換される。これを使って、次節以降の処理を行う。

```
/* path_define.sas */   *以下の①～⑤を指定して下さい。;
*① 符号表(Excel 形式)が保存されているパスの設定;
   %let folder = G:¥全消¥全消擬似ミクロデータ;
*② SAS データセット(ミクロデータ本体)の保存先 folder 名の指定;
*   ★この folder は①で指定した folder の直ぐ下に作成しておく★;
   %let ds_folder = zensho_parmanent;
*③ SAS データセット(ミクロデータ本体)のファイル名の指定;
*   ★このファイルは②で指定した folder に保存しておく★;
   %let dsname = original;
*④ ミクロデータ符号表(エクセルファイル)のファイル名の指定;
*   ★このファイルは①で指定した folder に保存しておくこと★;
   %let CodeTable = 統計センター_layout_code(鍵なし).xls;
*⑤ ミクロデータ符号表(エクセルファイル)のシート名の指定;
   %let sheet = sheet1;
*⑥ 統計調査名(日本語/英語)の指定; *途中結果表示の際に title 文で使用;
   %let SurveyName = 平成16年全国消費実態調査 擬似ミクロデータ;

libname xyz "&folder¥&ds_folder"; *この命令はそのまま;
```

プログラム1 プログラム実行環境の設定

```
/* giji_micro_zensho_codetable_import.sas */ *符号表をSASデータセットに変換;
proc import out=xyz.code_table
datafile="&folder¥&CodeTable" dbms=excel replace; getnames=yes; sheet="&sheet";
run;
proc print data=xyz.code_table(obs=56); title "&SurveyName.符号表"; run;
```

プログラム2 符号表をSASデータセットに変換

3.3 階層構造の符号表

符号表から変換したSASデータセットcode_table(付録３)は、いわゆる、階層構造型である。つまり、複数のオブザベーションでひとまとまりの情報を構成しており、SASで処理するのには若干の熟練度が必要となる。このSASデータセットをよく見ると、以下の規則性が読み取れる。

(1) 一つの変数に対する情報は、一行または複数行から成る。

(2) 変数の区切り目のオブザベーションのSAS変数VcodeContentsの値は「項目の区切り」

(3) 変数が数値型の場合は、変数Vtypeの値は「1」で、一つのオブザベーションから成る。

(4) 変数が文字型の場合は、変数vtypeの値は「欠損値」(半角ブランク)で複数のオブザベーションから成る。変数フォーマットが必要なのはこのタイプである。

(5) 変数フォーマット作成に必要な変数は、Vcode(コードの値)とVcodeContents(コードの意味)のペアである。

(6) 変数ラベル作成には、変数variableと変数Vlabelがあればできる。変数variableに変数名のローマ字表記のあるオブザベーションだけを対象にすればよい。

3.4 変数名の mnemonic 化と変数のラベル作成をする SAS プログラム:label_rename.sas

前節の(1)から(6)を考慮しながら、階層構造の SAS データセット code_table を読み取り、変数名の mnemonic 化と変数のラベル作成をするプログラムをプログラム３(次ページ)に示す。

このプログラムでは二つのテキストファイルを一旦 SAS の外に出力して、それらを再びこのプログラムの中で %include 文によって、命令文の一部として使っている。一つは、変数の rename に使用する右のテキストファイル F1toF14.txt、もう一つは、変数ラベルの作成のためのテキストファイル label.txt(付録 4)である。この二つのテキストファイルを見ると、誰でも進んでやりたくない手作業なので、自動化できることは作業の大幅な効率化になる。

このプログラムの各データステップで作成されるデータセットは、「結果ビューア」の①～⑤に表示しているので、プログラムの流れを理解し易い。

```
F1=SetaiKubun
F2=SetaiJinin
F3=ShuugyouJinin
F4=Kouzou
F5=Tatekata
F6=Shoyuu
F7=S1_Sex
F8=S1_Age
F9=S1_Shuugyou
F10=S1_KigyouKubun
F11=S1_KigyouKibo
F12=S1_Sangyou
F13=S1_Shokugyou
F14=Weight
```

ファイル名：
F1toF14.txt

3.5 変数 format を作成する SAS プログラム:format.sas

分析の過程や出力結果を見る場合、変数のコードのままでは大変読みにくいので、変数フォーマットを使用すべきである。変数の数が多いミクロデータの場合、この変数フォーマットを作成するのに手間暇がかかる。符号表の SAS データセットから、これを自動的に作成する SAS プログラム format.sas をプログラム４に示す。このプログラムを実行すると、SAS プログラム auto_format.sas (付録５)が自動作成される。

これを使用する際は、 %include 文で対象の SAS プログラムの中に埋め込めば良い。簡単な使用例を付録６に示す。

なお、プログラム１～４は http://mighty.gk.u-hyogo.ac.jp/confidential/Tutorial_Suoh2014.zip からダウンロード出来る。

4 まとめ

日本の政府機関が作成したミクロデータの提供を受けると、エクセル形式の「符号表」がメタデータとして付随してくる。変数名を分かり易い名前に rename をする、変数ラベルを付ける、変数フォーマットを定義するといった作業は、大変手間と時間がかかる。この作業は通常は手作業で行われるが、既に電子ファイル化された今回、符号表を SAS プログラムの入力データとして読み込み、こうした作業を全自動で行える SAS プログラムを開発した。今回は、平成 16 年全国消費実態調査の擬似ミクロデータの符号表を使用例として取り上げたが、他の政府統計の場合でも、ユーザーのパソコン環境の設定をする SAS プログラムを最初に実行するだけで、後のプログラムは変更なしに簡単に使えるように設計した。

SAS ユーザー総会のチュートリアルでは、これら一連の SAS プログラムの作成方法と解説を行う。階層構造型のファイルを SAS プログラムで処理するトレーニングにもなるはずである。これを通じて、SAS システムの持つ強力な「データハンドリング機能」を改めて認識する機械になれば、極めて有意義だと信じる。なお、本システムの利用方法は「利用の手引き」(付録２の後)を参照されたい。

参考文献

松井博(2008)『公的統計の体系と見方』日本評論社

```
/* label_rename.sas */ *教育用擬似ミクロデータの変数ラベル作成と変数の rename;

filename out1  "&folder¥label.txt";   *変数ラベル定義用テキスト(最後に自動削除);
filename out2  "&folder¥F1toF14.txt"; *変数 F1～F14 の rename 用テキスト(最後に自動削除);

* オリジナル変数 F15～F197 に変数ラベルを付すための LABEL 文の中味を外部ファイルに出力;
data label_define; keep var_label variable Vlabel;
    file out1; *LABEL 文の中味を外部ファイルに出力;
    length var_label $ 100;
    length No 8; *数値変数;
    set xyz.code_table;  ←――― 付録3
    if variable=" " then delete;

    if substr(variable,1,5)="Youto"
        then do; No=substr(variable,6,3);
                 var_label=compress('v' || No) || '="' || compress(Vlabel || '"');
             end;
        else var_label=compress(variable) || '="' || compress(Vlabel || '"');
    put var_label;
run;
proc print; title "label"; var var_label variable Vlabel; run;   ← 結果ビューア ①

data renameF1toF14; keep statement variable;
    file out2;
    *変数 F1～F14 を「符号表」にある変数名に変える rename 文の中味をテキストファイルに出力;
    array F F1-F14;
    set label_define;
    if _N_ <= 14 then do; statement=compress("F" || _N_ || "=" || variable);
                          put statement;
                      end;
                 else stop;
run;

proc print; title "renameF1toF14"; run;   ← 結果ビューア ②

data x; *変数 F1～F14 だけを rename をする;
    set xyz.&dsname;
    rename
    %include "&folder¥program¥出力テキストファイル¥F1toF14.txt";
    ;
proc print data=x(obs=40); title "変数 F1～F14 の rename"; run;   ← 結果ビューア ③

data x; drop i F15-F197; * オリジナル変数 F15～F197 を V1～V183 に rename する;
    set x;
    array V {183};
    array F {183} F15-F197;
    do i=1 to 183; V{i}=F{i}; end;
    label
    %include "&folder¥label.txt";
run;                                                    結果ビューア ④
                                                              ↓
proc print data=x(obs=20);       title "&SurveyName. (先頭 40 件:変数名表示)"; run;
proc print data=x(obs=20) label; title "&SurveyName. (先頭 40 件:変数ラベル表示)"; run;
                                                        結果ビューア ⑤ ↑
data xyz.x&dsname; set x; run; *ミクロデータのパーマネントデータセット完成;

x "del &folder¥label.txt"  ; x "echo label.txt を削除しました。";
x "del &folder¥F1toF14.txt"; x "echo F1toF14.txt を削除しました。";
```

プログラム3　変数名の mnemonic 化と変数ラベルの自動作成 SAS プログラム label_rename.sas

label

結果ビューア ①

OBS	var_label	variable	Vlabel
1	SetaiKubun="世帯区分"	SetaiKubun	世帯区分
2	SetaiJinin="世帯人員"	SetaiJinin	世帯人員
3	ShuugyouJinin="有業人員"	ShuugyouJinin	有業人員
4	Kouzou="住居の構造"	Kouzou	住居の構造
5	Tatekata="住居の建て方"	Tatekata	住居の建て方
6	Shoyuu="住居の所有関係"	Shoyuu	住居の所有関係
7	S1_Sex="性別"	S1_Sex	性別
8	S1_Age="年齢5歳階級"	S1_Age	年齢5歳階級
9	S1_Shuugyou="就業・非就業の別"	S1_Shuugyou	就業・非就業の別
10	S1_KigyouKubun="企業区分"	S1_KigyouKubun	企業区分
11	S1_KigyouKibo="企業規模"	S1_KigyouKibo	企業規模
12	S1_Sangyou="産業符号"	S1_Sangyou	産業符号
13	S1_Shokugyou="職業符号"	S1_Shokugyou	職業符号
14	Weight="集計用乗率"	Weight	集計用乗率
15	v1="年間収入"	Youto001	年間収入
16	v2="収入総額"	Youto002	収入総額
17	v3="実収入"	Youto003	実収入
18	v4="経常収入"	Youto004	経常収入
19	v5="勤め先収入"	Youto005	勤め先収入
		中略	
187	v173="個人・企業年金保険掛金"	Youto173	個人・企業年金保険掛金
188	v174="他の保険掛金"	Youto174	他の保険掛金
189	v175="有価証券購入"	Youto175	有価証券購入
190	v176="株式購入"	Youto176	株式購入
191	v177="他の有価証券購入"	Youto177	他の有価証券購入
192	v178="土地家屋借金返済"	Youto178	土地家屋借金返済
193	v179="他の借金返済"	Youto179	他の借金返済
194	v180="分割払・一括払購入借入金返済"	Youto180	分割払・一括払購入借入金返
195	v181="財産購入"	Youto181	財産購入
196	v182="その他"	Youto182	その他
197	v183="繰越金"	Youto183	繰越金

結果ビューア ②

renameF1toF14

OBS	variable	statement
1	SetaiKubun	F1=SetaiKubun
2	SetaiJinin	F2=SetaiJinin
3	ShuugyouJinin	F3=ShuugyouJinin
4	Kouzou	F4=Kouzou
5	Tatekata	F5=Tatekata
6	Shoyuu	F6=Shoyuu
7	S1_Sex	F7=S1_Sex
8	S1_Age	F8=S1_Age
9	S1_Shuugyou	F9=S1_Shuugyou
10	S1_KigyouKubun	F10=S1_KigyouKubun
11	S1_KigyouKibo	F11=S1_KigyouKibo
12	S1_Sangyou	F12=S1_Sangyou
13	S1_Shokugyou	F13=S1_Shokugyou
14	Weight	F14=Weight

変数F1～F14のrename

結果ビューア ③

OBS	SetaiKubun	SetaiJinin	ShuugyouJinin	Kouzou	Tatekata	Shoyuu	S1_Sex	S1_Age	S1_Shuugyou	S1_KigyouKubun	S1_KigyouKibo	S1_Sangyou	S1_Shokugyou	Weight	F15	F16	F17	F18	F19	F20
1	1																			
2	1																			
3	1																			

結果ビューア ④

平成16年全国消費実態調査 擬似ミクロデータ(先頭40件:変数名表示)

OBS	SetaiKubun	SetaiJinin	ShuugyouJinin	Kouzou	Tatekata	Shoyuu	S1_Sex	S1_Age	S1_ShuugyouKubun	S1_KigyouKubun	S1_KigyouKibo	S1_Sangyou	S1_Shokugyou	Weight	V1	V2	V3	V4	V5	V6	V7	V8	...中略...	V177	V178	V179	V180	V181	V182	V183
1																														
2																														
3																														

・・・以下略・・・

結果ビューア ⑤

平成16年全国消費実態調査 擬似ミクロデータ(先頭40件:変数ラベル表示)

OBS	世帯区分	世帯人員	有業人員	住居の構造	住居の建て方	住居の所有関係	性別	年齢5歳階級	就業・非就業の別	企業区分	企業規模	産業符号	職業符号	集計用乗率	年間収入	収入総額	実収入	経常収入	勤め先収入	事業・内職収入	農林漁業収入	家賃収入	...中略...	他の有価証券購入	土地家屋借金返済	他の借金返済	分割払・一括払購入借入金返済	財産購入	その他	繰越金
1																														
2																														
3																														

・・・以下略・・・

注 結果ビューアの④と⑤は対応している。

補足 システムのバージョンアップについて

プログラム2を少し書き換えれば、**利用の手引き**(付録２の下を参照)の「**プログラムの実行手順(1)**」を省略できることに気づいたが、本稿の原稿提出締め切り直前だったために、差し替え作業が間に合わなかった。当初、今回の教育用擬似ミクロデータだけに使うつもりで本システムを開発したが、汎用性を考慮すれば手順(1)はない方がはるかによい。

新バージョンでは、事前にユーザー側で、符号表のエクセルファイルのヘッダー行を「全て」削除し、中味のデータだけにしておく。これを proc import で読み込むと、変数名が自動的に F1, F2,・・・となる。その後に、data ステップを一つ挿入して、この変数 F1, F2,・・・を付録３にある変数名に rename するコードを挿入するだけで済む。チュートリアルの際は、この新バージョンを使って解説する。

なお、この**新プログラム2**は、3.5 節に記載の URL からダウンロードできるようにする。

```
/* format.sas */ *教育用擬似ミクロデータ用のproc format作成; options nocenter;

filename out1 "&folder¥auto_format.sas";

data format; keep formatName Vcode VcodeContents;
    set xyz.code_table; *符号表のデータセットを使用;
    retain sw formatName;

    if Vlabel="," then sw=0;
    if variable NE " " AND Vtype=" " /*文字型変数*/
        then do; sw=1; formatName=variable; output; return; end;
    if sw=1 then output;
run;

proc print; title "format対象変数(1)"; var formatName Vcode VcodeContents; run;

*★====== dataset 'format'に対して、「V」と「△」の処理を追加する======★;
data format; set format;
    if index(Vcode,"V") > 0 then Vcode=".";
    if (formatName="S1_KigyouKibo" OR formatName="S1_KigyouKubun") AND Vcode="△"
        then Vcode="99";
        else Vcode=kcompress(Vcode,"△");
run;
proc print; title "V converted to period"; run;
*★====== dataset 'format'に対する「V」と「△」の追加処理終了==========★;

proc sort data=format out=format1; by formatName; run;

data format1; set format1; by formatName;
    if first.formatName then statement2='value ' || compress('x' || formatName);
    statement=compress(Vcode || '=''' || VcodeContents || '''');
run;

proc print;
    title "format対象変数(2)";
    var formatName Vcode VcodeContents statement statement2;
run;

data _null_; file out1;
    set format1 end=final; by formatName;
    if _n_=1 then put "proc format;";
    if first.formatName then put statement2;
    if last.formatName  then put "    " statement ";" /;
                        else put "    " statement;
    if final then put "run;";
run;
```

プログラム4　SASフォーマットの自動作成プログラム:format.sas

付録1　平成16年全国消費実態調査擬似ミクロデータ符号表
(オリジナル先頭部)

政府統計コード		実施時期	平成16年	作成日	訂正日
統計調査名	全国消費実態	集計区分	2人以上勤労世帯	ファイル名	
調査票名	擬似ミクロデー	(備考・補足事項)			(照会先等)
コード体系	Shift_JIS	「△」はブランク(空白)を示す。			
レコード長	2974				

行番号	項目名	階層	位置	バイト数	配置	型	種別	変数名	符号	符号内容	備考
1	世帯区分	1	1	1				SetaiKubun	1	勤労	
2									2	勤労以外	
3									3	無職	
4	,	1	2	1					,	項目の区切り	
5	世帯票	1									
6	世帯人員	2	3	2	2	1		SetaiJinin	△2～	２人～	
7	,	2	5	1					,	項目の区切り	
8	有業人員	2	6	2	2	1		ShuugyouJin	△1～	１人～	
9									VV	不詳	
10	,	2	8	1					,	項目の区切り	

以下略(残りの内容は付録3と同じ)

付録2　平成16年全国消費実態調査擬似ミクロデータ符号表
(統計センター_layout_code(鍵なし).xls 先頭部)

lineNO	Vlabel	strata	position	bytes	xhaichi	Vtype	xshubetsu	variable	Vcode	VcodeContents
1	世帯区分	1	1	1				SetaiKubun	1	勤労
2									2	勤労以外
3									3	無職
4	,	1	2	1					,	項目の区切り
5	世帯票	1								
6	世帯人員	2	3	2	2	1		SetaiJinin	△2～	2人～
7	,	2	5	1					,	項目の区切り
8	有業人員	2	6	2	2	1		ShuugyouJinin	△1～	1人～
9									VV	不詳
10	,	2	8	1					,	項目の区切り

利用の手引き

最初にSASプログラム path_define.sas で以下の①～⑥を指定して下さい。

① 符号表(Excel 形式)が保存されているパスの設定
%let folder = G:¥全消¥全消擬似ミクロデータ ;
② SASデータセット(ミクロデータ本体)の保存先 folder 名の指定
★この folder は①で指定した folder の直ぐ下に作成しておく★
%let ds_folder = zensho_parmanent ;
③ SASデータセット(ミクロデータ本体)のファイル名の指定
★このファイルは②で指定した folder に保存しておく★
%let dsname = original ;
④ ミクロデータ符号表(エクセルファイル)のファイル名の指定
★このファイルは①で指定した folder に保存しておくこと★
%let CodeTable = 統計センター_layout_code(鍵なし).xls ;
⑤ ミクロデータ符号表(エクセルファイル)のシート名の指定
%let sheet = sheet1 ;
⑥ 統計調査名(日本語/英語)の指定;
%let SurveyName = 平成16年全国消費実態調査 擬似ミクロデータ ;

プログラムの実行手順

(1) 3.2 節の太字下線部で述べる前処理を符号表に対して行う。
(付録2の一行目参照)
この変数名がSASプログラムで使われる。

注
補足(結果ビューア⑤の下)を参照)

(2) 本文のプログラム①～④までを順番に実行する。

ファイル配置例　G:¥全消¥全消擬似ミクロデータ(①)¥統計センター_layout_code(鍵なし).xls(④)
G:¥全消¥全消擬似ミクロデータ¥zensho_parmanent(②)¥original.sas7bdat(③)

付録3　平成16年全国消費実態調査擬似ミクロデータ符号表

(SAS データセット code_table の冒頭部)

平成16年全国消費実態調査 擬似ミクロデータ符号表

OBS	lineNO	Vlabel	strata	position	bytes	xhaichi	Vtype	xshubetsu	variable	Vcode	VcodeContents
1	1	世帯区分	1	1	1	.			SetaiKubun	1	勤労
2	2		.	.	.	.				2	勤労以外
3	3		.	.	.	.				3	無職
4	4	,	1	2	1	.				,	項目の区切り
5	5	世帯票	1	.	.	.					
6	6	世帯人員	2	3	2	2	1		SetaiJinin	Δ2～	２人～
7	7	,	2	5	1	.				,	項目の区切り
8	8	有業人員	2	6	2	2	1		ShuugyouJinin	Δ1～	１人～
9	9		.	.	.	.				VV	不詳
10	10	,	2	8	1	.				,	項目の区切り
11	11	現住居等に関する事項	2	.	.	.					
12	12	住居の構造	3	9	1	.			Kouzou	1	木造
13	13		.	.	.	.				2	防火木造
14	14		.	.	.	.				3	鉄骨・鉄筋コンクリート造
15	15		.	.	.	.				4	その他（ブロック造り、レンガ造りなど）
16	16		.	.	.	.				V	不詳
17	17	,	3	10	1	.				,	項目の区切り
18	18	住居の建て方	3	11	1	.			Tatekata	1	一戸建
19	19		.	.	.	.				2	長屋建
20	20		.	.	.	.				3	共同住宅（１・２階建）
21	21		.	.	.	.				4	共同住宅（３～５階建）
22	22		.	.	.	.				5	共同住宅（６～１０階建）
23	23		.	.	.	.				6	共同住宅（１１階建以上）
24	24		.	.	.	.				7	その他
25	25		.	.	.	.				V	不詳
26	26	,	3	12	1	.				,	項目の区切り
27	27	住居の所有関係	3	13	1	.			Shoyuu	1	持ち家（世帯員名義）
28	28		.	.	.	.				2	持ち家（その他名義）
29	29		.	.	.	.				3	民営賃貸住宅（設備専用）
30	30		.	.	.	.				4	民営賃貸住宅（設備共用）
31	31		.	.	.	.				5	県市区町村営賃貸住宅
32	32		.	.	.	.				6	都市再生機構・公社等賃貸住宅
33	33		.	.	.	.				7	社宅・公務員住宅（借上げ含む）
34	34		.	.	.	.				8	借間
35	35		.	.	.	.				9	寮・寄宿舎
36	36		.	.	.	.				V	不詳
37	37	,	2	14	1	.				,	項目の区切り
38	38	世帯員に関する事項	2	.	.	.					
39	39	世帯主	3	.	.	.					
40	40	性別	4	15	1	.			S1_Sex	1	男
41	41		.	.	.	.				2	女
42	42	,	4	16	1	.				,	項目の区切り
43	43	年齢5歳階級	4	17	2	.			S1_Age	Δ1	２４歳未満
44	44		.	.	.	.				Δ2	２５～２９歳
45	45		.	.	.	.				Δ3	３０～３４歳
46	46		.	.	.	.				Δ4	３５～３９歳
47	47		.	.	.	.				Δ5	４０～４４歳
48	48		.	.	.	.				Δ6	４５～４９歳
49	49		.	.	.	.				Δ7	５０～５４歳
50	50		.	.	.	.				Δ8	５５～５９歳
51	51		.	.	.	.				Δ9	６０～６４歳
52	52		.	.	.	.				10	６５～６９歳
53	53		.	.	.	.				11	７０～７４歳
54	54		.	.	.	.				12	７５歳以上
55	55		.	.	.	.				VV	不詳
56	56	,	4	19	1	.				,	項目の区切り

SASデータセット CODE_TABLE：527オブザベーション、11変数

付録4　全国消費実態調査擬似ミクロデータ

(SAS 変数ラベルの定義用テキストファイル：label.txt)

```
SetaiKubun="世帯区分"
SetaiJinin="世帯人員"
ShuugyouJinin="有業人員"
Kouzou="住居の構造"
Tatekata="住居の建て方"
Shoyuu="住居の所有関係"
S1_Sex="性別"
S1_Age="年齢5歳階級"
S1_Shuugyou="就業・非就業の別"
S1_KigyouKubun="企業区分"
S1_KigyouKibo="企業規模"
S1_Sangyou="産業符号"
S1_Shokugyou="職業符号"
Weight="集計用乗率"
v1="年間収入"
v2="収入総額"
v3="実収入"
v4="経常収入"
v5="勤め先収入"
v6="事業・内職収入"
v7="農林漁業収入"
v8="家賃収入"
v9="他の事業収入"
v10="内職収入"
v11="本業以外の勤め先・事業・内職収入"
v12="他の経常収入"
v13="財産収入"
v14="社会保障給付"
v15="公的年金給付"
v16="他の社会保障給付"
v17="仕送り金"
v18="特別収入"
v19="受贈金"
v20="その他"
v21="実収入以外の収入"
v22="預貯金引出"
v23="保険取金"
v24="個人・企業年金保険取金"
v25="他の保険取金"
v26="有価証券売却"
v27="株式売却"
v28="他の有価証券売却"
v29="土地家屋借入金"
v30="他の借入金"
v31="分割払・一括払購入借入金"
v32="財産売却"
v33="その他"
v34="繰入金"
v35="支出総額"
v36="実支出"
v37="消費支出"
v38="食料"
v39="穀類"
v40="米"
v41="パン"
v42="めん類"
v43="他の穀類"
v44="魚介類"
v45="生鮮魚介"
v46="塩干魚介"
v47="魚肉練製品"
v48="他の魚介加工品"
v49="肉類"
v50="生鮮肉"
v51="加工肉"
v52="乳卵類"
v53="牛乳"
v54="乳製品"
v55="卵"
v56="野菜・海藻"
v57="生鮮野菜"
v58="乾物・海藻"
v59="大豆加工品"
v60="他の野菜・海藻加工品"
v61="果物"
v62="生鮮果物"
v63="果物加工品"
v64="油脂・調味料"
v65="油脂"
v66="調味料"
v67="菓子類"
v68="調理食品"
v69="主食的調理食品"
v70="他の調理食品"
v71="飲料"
v72="茶類"
v73="コーヒー・ココア"
v74="他の飲料"
v75="酒類"
v76="外食"
v77="一般外食"
v78="学校給食"
v79="住居"
v80="家賃地代"
v81="設備修繕・維持"
v82="設備材料"
v83="工事その他のサービス"
v84="光熱・水道"
v85="電気代"
v86="ガス代"
v87="他の光熱"
v88="上下水道料"
v89="家具・家事用品"
v90="家庭用耐久財"
v91="家事用耐久財"
v92="冷暖房用器具"
v93="一般家具"
v94="室内装備・装飾品"
v95="寝具類"
v96="家事雑貨"
v97="家事用消耗品"
v98="家事サービス"
v99="被服及び履物"
v100="和服"
v101="洋服"
v102="男子用洋服"
v103="婦人用洋服"
v104="子供用洋服"
v105="シャツ・セーター類"
v106="男子用シャツ・セーター類"
v107="婦人用シャツ・セーター類"
v108="子供用シャツ・セーター類"
v109="下着類"
v110="男子用下着類"
v111="婦人用下着類"
v112="子供用下着類"
v113="生地・糸類"
v114="他の被服"
v115="履物類"
v116="被服関連サービス"
v117="保健医療"
v118="医薬品"
v119="健康保持用摂取品"
v120="保健医療用品・器具"
v121="保健医療サービス"
v122="交通・通信"
v123="交通"
v124="自動車等関係費"
v125="自動車等購入"
v126="自転車購入"
v127="自動車等維持"
v128="通信"
v129="教育"
v130="授業料等"
v131="教科書・学習参考教材"
v132="補習教育"
v133="教養娯楽"
v134="教養娯楽用耐久財"
v135="教養娯楽用品"
v136="書籍・他の印刷物"
v137="教養娯楽サービス"
v138="宿泊料"
v139="パック旅行費"
v140="月謝類"
v141="他の教養娯楽サービス"
v142="その他の消費支出"
v143="諸雑費"
v144="理美容サービス"
v145="理美容用品"
v146="身の回り用品"
v147="たばこ"
v148="その他の諸雑費"
v149="こづかい(使途不明)"
v150="交際費"
v151="食料"
v152="家具・家事用品"
v153="被服及び履物"
v154="教養娯楽"
v155="他の物品サービス"
v156="贈与金"
v157="他の交際費"
v158="仕送り金"
v159="非消費支出"
v160="直接税"
v161="勤労所得税"
v162="個人住民税"
v163="他の税"
v164="社会保険料"
v165="公的年金保険料"
v166="健康保険料"
v167="介護保険料"
v168="他の社会保険料"
v169="他の非消費支出"
v170="実支出以外の支出"
v171="預貯金"
v172="保険掛金"
v173="個人・企業年金保険掛金"
v174="他の保険掛金"
v175="有価証券購入"
v176="株式購入"
v177="他の有価証券購入"
v178="土地家屋借金返済"
v179="他の借金返済"
v180="分割払・一括払購入借入金返済"
v181="財産購入"
v182="その他"
v183="繰越金"
```

付録5　全国消費実態調査擬似ミクロデータ(SAS 変数用フォーマット)

auto_format.sas

```
proc format;
value xKouzou
    1="木造"
    2="防火木造"
    3="鉄骨・鉄筋コンクリート造"
    4="その他(ブロック造り、レンガ造りなど)"
    .="不詳" ;

value xS1_Age
    1="24歳未満"
    2="25～29歳"
    3="30～34歳"
    4="35～39歳"
    5="40～44歳"
    6="45～49歳"
    7="50～54歳"
    8="55～59歳"
    9="60～64歳"
    10="65～69歳"
    11="70～74歳"
    12="75歳以上"
    .="不詳" ;

value xS1_KigyouKibo
    1="1～4人"
    2="5～29人"
    3="30～499人"
    4="500～999人"
    5="1000人以上"
    99="非就業又は官公"
    .="不詳" ;

value xS1_KigyouKubun
    1="民営"
    2="自営"
    3="官公"
    99="非就業"
    .="不詳" ;

value xS1_Sangyou
    1="農業"
    2="林業"
    3="漁業"
    4="鉱業"
    5="建設業"
    6="製造業"
    7="電気・ガス・熱供給・水道業"
    8="情報通信業"
    9="運輸業"
    10="卸売・小売業"
    11="金融・保険業"
    12="不動産業"
    13="飲食店・宿泊業"
    14="医療・福祉"
    15="教育・学習支援業"
    16="複合サービス事業"
    17="サービス業(他に分類されないもの)"
    18="公務(他に分類されないもの)"
    19="その他(非就業を含む)"
    .="不詳" ;

value xS1_Sex
    1="男"
    2="女" ;

value xS1_Shokugyou
    1="常用労務作業者"
    2="臨時及び日々雇労務作業者"
    3="民間職員"
    4="官公職員1"
    5="官公職員2"
    6="商人及び職人"
    7="個人経営者"
    8="農林漁業従事者"
    9="法人経営者"
    10="自由業者"
    11="その他"
    12="無職"
    .="不詳" ;

value xS1_Shuugyou
    1="就業"
    2="  うちパート"
    3="非就業"
    4="  うち仕事を探している"
    .="不詳" ;

value xSetaiKubun
    1="勤労"
    2="勤労以外"
    3="無職" ;

value xShoyuu
    1="持ち家(世帯員名義)"
    2="持ち家(その他名義)"
    3="民営賃貸住宅(設備専用)"
    4="民営賃貸住宅(設備共用)"
    5="県市区町村営賃貸住宅"
    6="都市再生機構・公社等賃貸住宅"
    7="社宅・公務員住宅(借上げ含む)"
    8="借間"
    9="寮・寄宿舎"
    .="不詳" ;

value xTatekata
    1="一戸建"
    2="長屋建"
    3="共同住宅(1・2階建)"
    4="共同住宅(3～5階建)"
    5="共同住宅(6～10階建)"
    6="共同住宅(11階建以上)"
    7="その他"
    .="不詳" ;

run;
```

付録6　SAS 変数用フォーマットプログラム auto_format.sas の使用例

```
/* freq.sas */ options nocenter;
*擬似ミクロデータ用 SAS フォーマット設定プログラム「auto_format.sas」の使用例;
%let microdata=xoriginal; * ユーザーが付けたミクロデータ本体(SAS data set)のファイル名;
libname micro "G:¥全消¥全消擬似ミクロデータ¥zensho_parmanent"; * ユーザーの環境に設定;
%include "G:¥全消¥全消擬似ミクロデータ¥auto_format.sas"; *proc format プログラム実行;

title "擬似ミクロデータ 度数分布表";

%macro freq(variable);
proc freq data=micro.xoriginal; tables &variable / missing;
    format &variable x&variable..;
run;
%mend;

%freq(SetaiKubun)
%freq(Kouzou)
%freq(Tatekata)
%freq(Shoyuu)
%freq(S1_Sex)
%freq(S1_Age)
%freq(S1_Shuugyou)
%freq(S1_KigyouKubun)
%freq(S1_KigyouKibo)
%freq(S1_Sangyou)
%freq(S1_Shokugyou)
```

このプログラムは平成 16 年全国消費実態調査の擬似ミクロデータに含まれる全ての分類変数の度数分布表を出力する。%include 文によって、実行時に全ての変数のフォーマットを定義する auto_format.sas が実行される。その結果、マクロ freq 内の format 文が有効になり、出力されるクロス表の表側には、変数コードではなくて、下に示すように、format 値が表示される。

なお、表頭には変数名の代わりに変数ラベルが表示されている。

実行結果のうち、2 番目から 5 番目までの 4 つの表「結果ビューア」画面を以下に示す。

擬似ミクロデータ 度数分布表

FREQ プロシジャ

住居の構造

Kouzou	度数	パーセント	累積度数	累積パーセント
不詳	5584	17.43	5584	17.43
木造	14064	43.91	19648	61.35
防火木造	3626	11.32	23274	72.67
鉄骨・鉄筋コンクリート造	8739	27.29	32013	99.95
その他(ブロック造り、レンガ造りなど)	15	0.05	32028	100.00

擬似ミクロデータ 度数分布表

FREQ プロシジャ

住居の所有関係

Shoyuu	度数	パーセント	累積度数	累積パーセント
不詳	1446	4.51	1446	4.51
持ち家(世帯員名義)	22936	71.61	24382	76.13
持ち家(その他名義)	785	2.45	25167	78.58
民営賃貸住宅(設備専用)	3952	12.34	29119	90.92
民営賃貸住宅(設備共用)	3	0.01	29122	90.93
県市区町村営賃貸住宅	1287	4.02	30409	94.95
都市再生機構・公社等賃貸住宅	325	1.01	30734	95.96
社宅・公務員住宅(借上げ含む)	1282	4.00	32016	99.96
借間	12	0.04	32028	100.00

擬似ミクロデータ 度数分布表

FREQ プロシジャ

住居の建て方

Tatekata	度数	パーセント	累積度数	累積パーセント
不詳	5665	17.69	5665	17.69
一戸建	21623	67.51	27288	85.20
長屋建	87	0.27	27375	85.47
共同住宅(1・2階建)	1055	3.29	28430	88.77
共同住宅(3～5階建)	2679	8.36	31109	97.13
共同住宅(6～10階建)	641	2.00	31750	99.13
共同住宅(11階建以上)	278	0.87	32028	100.00

擬似ミクロデータ 度数分布表

FREQ プロシジャ

性別

S1_Sex	度数	パーセント	累積度数	累積パーセント
.	1	0.00	1	0.00
男	29381	91.74	29382	91.74
女	2645	8.26	32028	100.00

Let's データ分析

Let'sデータ分析ミクロデータ分析コンテスト：規定課題

宇野　慧
アステラス製薬株式会社　開発本部　データサイエンス部

要旨

本稿では、平成16年度全国消費実態調査の個票データに基づき、独立行政法人統計センターが作成した教育用擬似ミクロデータを用いて、コンテストの規定課題再現の方法及び再現結果を示した。

キーワード：疑似ミクロデータ、TABULATEプロシジャ

1. 用いたデータ

本稿で用いたデータは平成16年度全国消費実態調査の個票データに基づき、独立行政法人統計センターが作成した教育用擬似ミクロデータである。データは32027行、183変数であり、以下ではデータセット名DATAとする。

2. 表1-1、表1-2、表1-3の再現

表1-1　集計世帯表（各レコードを単純にカウントした表）、表1-2　世帯数分布（各レコードを集計用乗率で重み付けしてカウントした表）、表1-3　世帯数分布（世帯数分布を10万分比でカウントした表）の3表について、以下のプログラムを実行することで規定課題の結果を得ることが出来る。

```
/*初めに、課題表の出力のために必要系列の作成を行う*/
DATA data; SET data;
Weight2 = Weight/4.95465;  /*10万人で基準化した度数分布表を作成するために乗数を作成する*/
INFORMAT d_12 $1.;
 IF Shuugyoujinin in (1,2) then d_12 = "1"; /*就業人員 1or2名世帯のダミーを作成する*/
 ELSE d_12 = "2";RUN;

PROC FORMAT;
 VALUE Setaijininf 1= '1人' 2= '2人' 3= '3人' 4= '4人' 5= '5人' 6= '6人' 7= '7人' 8= '8人' 9= '9人' 10= '10人';
 VALUE Shuugyoujininf 1 = '1人' 2 = '2人' 3 = '3人' 4 = '4人' 5 = '5人' 6 = '6人' 9999 = '不詳';
 VALUE $d_12f 1="（特掲）1人又は2人" 2="3名以上";RUN;

/*①の世帯数集計*/
PROC TABULATE DATA=data MISSING; CLASS ShuugyouJinin SetaiJinin;
 FORMAT Setaijinin Setaijininf. Shuugyoujinin Shuugyoujininf.;
 KEYLABEL N=' ';   /*keylabelで"N"を出力しないようにする*/
 TABLES ShuugyouJinin='就業人員 ' all = '合計', SetaiJinin =' 世帯人員' all = '合計'
 /BOX = '表1-1 世帯数' MISSTEXT='0';RUN;
```

```
/*②の集計用乗率で重み付けした集計*/
PROC TABULATE DATA=data MISSING; CLASS ShuugyouJinin SetaiJinin;
 FORMAT Setaijinin Setaijininf. Shuugyoujinin Shuugyoujininf.;
 VAR / WEIGHT=weight;  /*集計乗数をウェイトとして設定*/  KEYLABEL N=' ' SUM=' ';
 TABLES ShuugyouJinin = '就業人員' * Weight="" * (SUM*F=comma8.)
        all='合計' * Weight="" * (SUM*F=comma8.) ,  (SetaiJinin ='  世帯人員' all = '合計')*(f=comma8.)
 /BOX = '表 1-2  重み付け世帯数' MISSTEXT='0';RUN;

/*③の 10 万分比で基準化した集計*/
PROC TABULATE DATA=data MISSING; CLASS ShuugyouJinin SetaiJinin;
 FORMAT Setaijinin Setaijininf. Shuugyoujinin Shuugyoujininf.;
 VAR Weight2;  /*10 万分比のウェイト*  KEYLABEL N=' ' SUM=' ';
 TABLES  ShuugyouJinin = '就業人員' * Weight2="" * (SUM*F=comma8.)
         all='合計' * Weight2="" * (SUM*F=comma8.) ,  (SetaiJinin ='  世帯人員' all = '合計')*(f=comma8.)
 /BOX = '表 1-3 10 万分比基準化世帯数' MISSTEXT='0';RUN;
```

3. 表 2 の再現

表 2　支出額（消費支出及び 10 大費目）について、以下のプログラムを実行することで規定課題の結果を得ることが出来る。

```
/*④4 人世帯の支出額の集計値*/
PROC TABULATE DATA=data MISSING; CLASS ShuugyouJinin;WEIGHT Weight;
 FORMAT Setaijinin Setaijininf. Shuugyoujinin Shuugyoujininf.;
 VAR Youto037 Youto038 Youto079 Youto084 Youto089 Youto099
     Youto117 Youto122 Youto129 Youto133 Youto142;
 TABLES ALL="総数", N="集計世帯数"*f=comma8.
    Youto037=""*(SUMWGT="重み付け世帯数"*f=comma8.)
   (Youto037="食費支出" Youto038="食料" Youto079="住居"  Youto084="光熱・水道"
    Youto089="家具・家事用品" Youto099="被服及び履物" Youto117="保健医療" Youto122="交通・通信"
    Youto129="教育" Youto133="教養娯楽" Youto142="その他消費支出")
    *(MEAN=""*f=comma8.)/ MISSTEXT = "0";RUN;

/*⑤4 人世帯のうち、就業人員 1 人または 2 人世帯の支出額の集計値*/
PROC TABULATE DATA=data MISSING; CLASS ShuugyouJinin d_12; WEIGHT Weight;
 FORMAT Setaijinin Setaijininf. Shuugyoujinin Shuugyoujininf. d_12 $d_12f.;
 VAR Youto037 Youto038 Youto079 Youto084 Youto089 Youto099
     Youto117 Youto122 Youto129 Youto133 Youto142;
 TABLE ALL="4 人世帯合計" ShuugyouJinin="就業人員数" d_12="", N="集計世帯数"*f=comma8.
   Youto037=""*(SUMWGT="重み付け世帯数"*f=comma8.)
  (Youto037="食費支出" Youto038="食料" Youto079="住居" Youto084="光熱・水道" Youto089="家具・家事用品"
 Youto099="被服及び履物" Youto117="保健医療" Youto122="交通・通信" Youto129="教育" Youto133="教養娯楽"
 Youto142="その他消費支出")*(MEAN=""*f=comma8.)/ MISSTEXT = "0"   WHERE Setaijinin = 4;RUN;
```

表 1-1 集計世帯表（各レコードを単純にカウントした表）

表 1-1	世帯人員									合計
世帯数	2人	3人	4人	5人	6人	7人	8人	9人	10人	
就業人員										
1人	4,124	3,908	4,132	1,436	256	51	6	0	0	13,913
2人	3,239	3,391	4,201	1,943	494	162	29	0	0	13,459
3人	0	1,035	1,031	559	232	84	6	3	0	2,950
4人	0	0	324	220	104	31	12	0	0	691
5人	0	0	0	27	6	7	0	0	0	40
6人	0	0	0	0	3	3	0	0	0	6
不詳	75	203	256	220	119	52	28	12	3	968
合計	7,438	8,537	9,944	4,405	1,214	390	81	15	3	32,027

表 1-2　世帯数分布（各レコードを集計用乗率で重み付けしてカウントした表）

表 1-2 重み付	世帯人員									合計
1)世帯数	2人	3人	4人	5人	6人	7人	8人	9人	10人	
就業人員										
1人	64,691	61,284	66,299	22,740	3,813	831	84	0	0	219,743
2人	50,138	51,523	64,868	29,616	6,801	2,336	441	0	0	205,723
3人	0	15,912	15,615	7,963	3,302	1,137	76	36	0	44,041
4人	0	0	4,851	3,345	1,354	422	195	0	0	10,168
5人	0	0	0	383	80	108	0	0	0	570
6人	0	0	0	0	49	51	0	0	0	99
不詳	1,006	3,287	4,216	3,519	1,761	727	401	170	33	15,120
合計	115,835	132,005	155,850	67,565	17,161	5,611	1,197	207	33	495,465

表 1-3　世帯数分布（世帯数分布を 10 万分比でカウントした表）

表 1-3 10万分比基	世帯人員									合計
準化世帯数	2人	3人	4人	5人	6人	7人	8人	9人	10人	
就業人員										
1人	13,057	12,369	13,381	4,590	770	168	17	0	0	44,351
2人	10,119	10,399	13,092	5,977	1,373	471	89	0	0	41,521
3人	0	3,211	3,152	1,607	666	229	15	7	0	8,889
4人	0	0	979	675	273	85	39	0	0	2,052
5人	0	0	0	77	16	22	0	0	0	115
6人	0	0	0	0	10	10	0	0	0	20
不詳	203	664	851	710	355	147	81	34	7	3,052
合計	23,379	26,643	31,455	13,637	3,464	1,132	242	42	7	100,000

表２　支出額（消費支出及び10大費目）

総数	32,027	495,465	328,140	72,883	17,687	19,238	9,204	14,138	11,366	47,961	22,270	31,389	82,003
4人家族合計	9,944	155,850	335,438	76,362	15,345	20,214	8,885	14,452	10,987	47,894	33,442	32,269	75,588
世帯人員数													
1人	4,132	66,299	305,234	71,543	17,556	18,854	8,383	13,579	11,656	42,703	31,202	31,959	57,801
2人	4,201	64,868	347,740	78,472	12,932	20,621	8,917	14,876	10,538	52,141	39,485	33,681	76,078
3人	1,031	15,615	380,521	83,796	12,583	23,045	10,276	16,561	10,331	52,406	22,513	27,852	121,160
4人	324	4,851	399,962	85,083	18,500	22,490	11,763	14,096	10,812	51,310	4,509	32,769	148,628
不詳	256	4,216	379,882	82,134	24,296	22,238	7,843	14,238	10,011	43,521	49,453	31,206	94,942
1人又は2人	8,333	131,168	326,256	74,970	15,269	19,728	8,647	14,221	11,103	47,371	35,298	32,810	66,839
3人以上	1,611	24,682	384,233	83,765	15,747	22,798	10,153	15,680	10,371	50,673	23,576	29,391	122,080

同時方程式モデルを用いた健康/不健康支出の分析

宇野　慧
アステラス製薬株式会社　開発本部　データサイエンス部

要旨

保険に対する需要は、世帯のリスク選好度に関する一種の指標と考えることが出来る(宇野(2013))。本稿では擬似ミクロデータを用いて、生鮮野菜・生鮮果物といった健康支出、酒・タバコといった不健康支出に対する、貯蓄性の保険料支払い額の傾向を考察した。保険料支払い額自体も世帯の支出選択行動である点を考慮して、①保険料支払い額の推定式、②健康支出の推定式、③不健康支出の推定式で構成される同時方程式モデルを最尤法で推定した。その結果、保険料支払い額が多い世帯では、健康支出、不健康支出の両方が高い傾向が確認できた。また世帯属性のうち、世帯主年齢が不健康支出に対して強いコホート効果として影響している可能性が示唆された。

キーワード：擬似ミクロデータ、消費分析、同時方程式モデル、NLMIXED プロシジャ

1. はじめに

宇野(2013)では、世帯主の就業状況などの世帯属性が貯蓄性保険需要に対して与える影響を分析した。その結果として、公務員世帯の貯蓄性保険料支払い額が他と比較して高額であったことなどから、貯蓄性保険料支払い額が世帯のリスク選好度に関する指標となる可能性が示唆された。この結果を受け本稿では、貯蓄性保険料支払い額を用いて、リスク選好度が世帯の消費行動においても観察できるかどうかを検証した。特に、世帯のリスク選好度が顕著に反映されると考えられる項目として、生鮮野菜・生鮮果物といった健康関連支出、および、酒やタバコといった不健康関連支出に着目した。本稿で用いたデータは、平成 16 年度全国消費実態調査から作成された教育用疑似ミクロデータ(2 人以上の勤労者世帯；32037 世帯)である。

以下で本稿の構成を概説する。2 節では分析に用いたデータの変数名や作成方法などについて説明する。また健康支出、不健康支出それぞれについて、基本統計量と度数分布図により分布を確認する。加えて保険料支払い額との散布図を両変数について求め、変数間の大まかな関係を把握する。3 節では本稿で行う分析に関する理論的な背景を説明すると共に、推定に用いたプログラムの概要を記載する。4 節では推定結果の概略を説明し、加えて簡単な解釈を行う。5 節では本稿のまとめと、今後の分析に対する展望を述べる。

2. データセットの特性について

2-1. 分析に用いた変数名、および作成方法

本稿の分析で用いた変数の情報について、以下の表 2-1 にまとめた。式表記に関しては、3 節のモデル設定の箇所で参照するために表記している。

表 2-1　分析に用いた変数一覧

変数名	式表記	詳細説明
保険料支出	y_1	貯蓄性保険料の支払い額(Youto174)を自然対数変換した値　※支出額が 0 の場合は、そのまま 0 としているが、全体の 1%程度であり分析には大きく影響しない
健康支出	y_2	生鮮野菜(Youto057)と生鮮果物(Youto062)の支出額を足し、自然対数変換した値
不健康支出	y_3	酒(Youto075)とタバコ(Youto147) の支出額を足し、自然対数変換した値　※支出額が 0 の場合は、そのまま 0 としているが、全体の 1%未満であり分析には大きく影響しない
大企業ダミー	x	世帯主の就業先規模(S1_KigyouKibo)が 4：500～999 人または 5：1000 人以上の場合に 1 を取るダミー変数　※就業属性については、中小企業に就業している世帯を基準に設定
公務員ダミー		世帯主の就業先区分(S1_KigyouKubun)が 3：官公の場合に 1 を取るダミー変数 ※就業属性については、中小企業に就業している世帯を基準に設定
経常所得		経常所得(Youto004)を自然対数変換した値
女性ダミー		世帯主性別(S1_Sex)が 2 の場合に 1 をとるダミー変数
年代ダミー		世帯主年齢(S1_Age)について、20 代～60 代以上のダミー変数を作成。推定での基準：20 代
世帯人数ダミー		世帯人員(SetaiJinin)について、2 人～6 人以上のダミー変数を作成。推定での基準：2 人
持家ダミー	z	住居の所有区分(Shoyuu)が 1(持家(世帯人名義))あるいは 2(持家(その他))の場合には 1 を取るダミー変数
住宅ローン完済ダミー		持家世帯、かつ住宅ローンの支払額(Youto178)がゼロの場合に 1 を取るダミー変数

2-2. 基本統計量、散布図

2-1 で設定した y_2:健康支出、y_3:不健康支出それぞれの基本統計量及び度数分布図を図 2-2-1、図 2-2-2 に示す。また、y_1:保険料との散布図をそれぞれ図 2-2-3、図 2-2-4 に示す。

図 2-2-1　健康支出の度数分布図、基本統計量

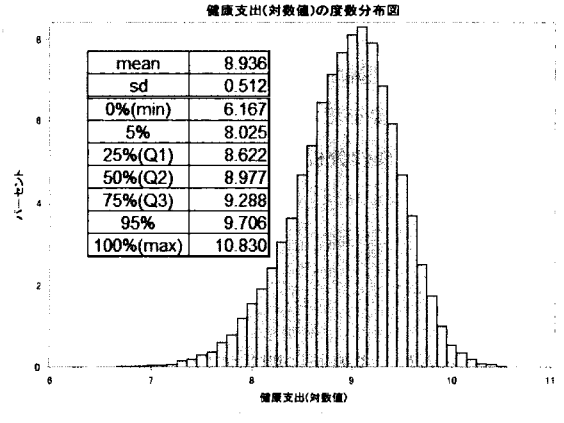
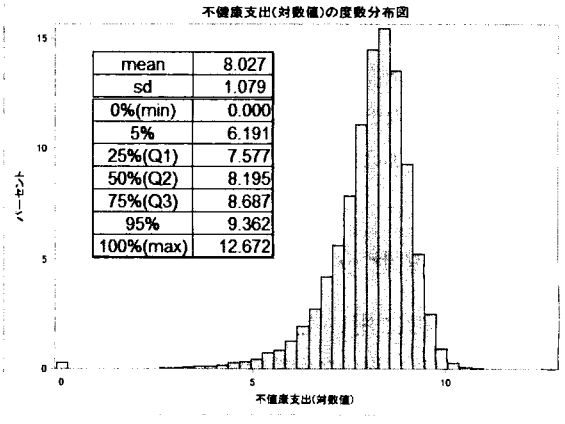


図 2-2-2　不健康支出の度数分布図、基本統計量



度数分布図、及び基本統計量から、健康支出に関しては対数変換により正規分布に近い形となる。一方で、不健康支出に関しては、対数変換を行っても左に裾を引いた分布となることが分かる。また、不健康支出に関しては支出額がゼロの世帯がいるが、その割合は全体の１％未満である。

図 2-2-3　保険料と健康支出の散布図

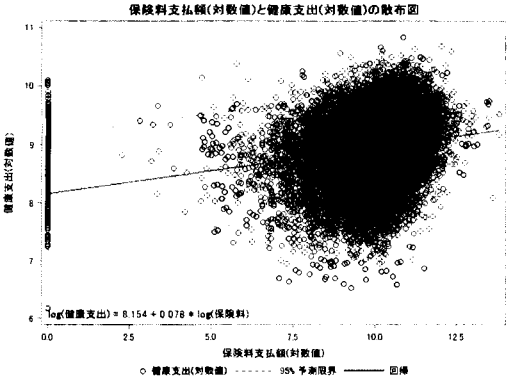
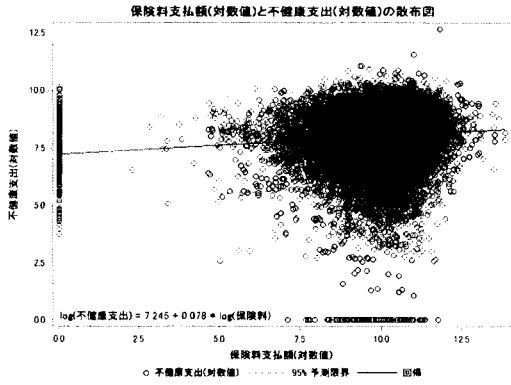


図 2-2-4　保険料と不健康支出の散布図



図からわかる通り、健康支出、不健康支出ともに保険料支払い額と弱い正の相関が確認出来る。単回帰を行った結果、保険料支払い額の係数はどちらも 0.078 となり、保険料支払い額が高い世帯ほど、健康支出・不健康支出の額がやや高まる傾向が見られた。

次に、分析に用いる各世帯属性と、健康/不健康支出との関係を確認する。文中で言及している図表については、本文末尾の補足に掲載している。なお、各世帯属性には不詳データが含まれることから、表によっては周辺度数を全て足し合わせても全世帯数(32037 世帯)に一致しない点に注意が必要である。

はじめに、健康支出に関して検討する。図 2-2-5 から分かる通り、モデル推定で基準とする中小企業に比べ、大企業と公務員では高い支出水準であることが分かる。また、図 2-2-6 から分かる通り、世帯主が女性の世帯では支出額の水準が低い。世帯主の年代に関しては図 2-2-7 から分かる通り、年代が上がるにつれて支出額も単調に増加している。最後に、図 2-2-8 から世帯人員の増加につれ、支出額も単調増加している事が確認できる。

続いて、不健康支出に関して検討する。図 2-2-5 から、中小企業、公務員と比べ大企業でやや低い傾向が見られる。また、図 2-2-6 から女性世帯では支出額が低いことが分かる。図 2-2-7 では、モデル推定で基準となる 30 代の支出が最も低く、また 50 代と 60 代以上ではあまり差がないことが分かる。最後に世帯人員数に関して図 2-2-8 からは、6 人以上の大規模世帯では支出額が多い傾向があるものの、全体的にあまり大きな違いが無いことが確認できる。

次節では同時方程式モデルを構築し、保険料支払い額以外の世帯属性などを考慮することにより、本節で検討した支出の傾向について考察する。

3. 推定モデル

本節では、推定モデルの構成を解説し、最尤推定を行う尤度式を示す。本稿で分析する同時方程式モデルは、①貯蓄性保険需要に関する推定式、②健康支出に関する推定式、③不健康支出に関する推定式の 3 式で構成される。貯蓄性保険需要はそれ自体が世帯の決定変数であると同時に、世帯のリスク選好度の指標とし

て健康支出、不健康支出に関係すると想定し、①式の被説明変数、および②式・③式の説明変数としてモデルに盛り込んだ。

モデルを構成するその他の説明変数として、世帯主属性(就業状況ダミー、年代ダミー、性別ダミー)、世帯属性(人員数ダミー)、収入等(経常所得)を3式に共通して用いた。また、①式については住居等に関係する属性として、持家ダミーと住宅ローン完済ダミーを追加でモデルに含めた。なお、宇野(2013)では保険料支払い額の推定に関して、0(無保険)を考慮した打切り分布の推定(Tobit モデルあるいは Hurdle モデル推定)を行った。しかしながら、無保険世帯が全体の1%程度であり推定結果に大きく影響しないことから、今回の分析では打切り分布は考慮しなかった。

以上の設定を数式で記述すると、以下のようになる。

$$\begin{cases} y_1 = x\beta_1 + z\gamma_1 + \varepsilon_1 \\ y_2 = x\beta_2 + y_1\delta_2 + \varepsilon_2 \\ y_3 = x\beta_3 + y_1\delta_3 + \varepsilon_3 \end{cases}$$

行列表記を用いて、次のように誘導形に書くことが出来る。なお、推定の誤差項は、以下のような多変量正規分布に従うと仮定する。

$$\mathrm{Y} = \mathrm{XB} + \varepsilon$$

$$\Leftrightarrow \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} x & z \\ x & y_1 \\ x & y_1 \end{pmatrix} \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 \\ \gamma_1 & \delta_2 & \delta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} \quad \text{where} \quad \varepsilon \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix} \right)$$

また、推定式を以下のような構造形に変換して、Γ行列を求める。

$$\Gamma\mathrm{Y} + \mathrm{Z}\Lambda = \varepsilon, \qquad \Gamma = \begin{pmatrix} 1 & 0 & 0 \\ -\delta_2 & 1 & 0 \\ -\delta_3 & 0 & 1 \end{pmatrix}, \qquad \mathrm{Z}\Lambda = \begin{pmatrix} x & z \\ x & 0 \\ x & 0 \end{pmatrix} \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 \\ \gamma_1 & 0 & 0 \end{pmatrix}$$

以上の設定をもとに、以下の尤度関数が構成でき、この式をもとにパラメータの最尤推定値を求める。本稿は完全情報最尤法(FIML)に基づく尤度関数を定義している。しかしながら、保険料支払い額は頻繁に変更されることが無いことから、y_1 は y_2 や y_3 よりも前に決定されると考えられる。そのため y_2 や y_3 の推定式の説明変数に y_1 が入る一方で、y_1 の推定式の説明変数に y_2 や y_3 は含めない。また、y_2 と y_3 についても誤差の相関以外の関係は設定してないため、Γ行列の行列式は1となり推定には影響しない。

$$\log L(B,\Gamma,\Sigma) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\left(\left|\Sigma\right|^{-1}\right) + \log\left(\left|\Gamma\right|\right) - \frac{1}{2}\left(\mathrm{Y} - \mathrm{XB}\right)^T \Sigma^{-1}\left(\mathrm{Y} - \mathrm{XB}\right)$$

また、推定に先立ちパラメータの初期値を指定する必要がある。これについては、各推定式単体で最尤推定を行って得たパラメータ推定値を、初期値として設定した。

推定プログラムの概要は、以下の通りである。詳細については、別途プログラム本体を参照されたい。

```
proc nlmixed data=data tech = NEWRAP;
 /*パラメータの初期値設定*/
 parms b100 - b312 d2 d3 sigma11 - sigma33;

 /*誘導系の右辺を定義する*/
 xbeta = b100 + Daikigyou*b101 +… + NoLoan*b114 ;
 xgamma =  b200 + Daikigyou*b201 + …+ SetaiJinin6o*b213 + log_hoken*d2;
 xdelta = b300 + Daikigyou*b301 + …+SetaiJinin6o*b313 + log_hoken*d3;

 /*誘導系のモデル式を定義する*/
 e1 = log_hoken - xbeta;  e2 = log_yasai1 - xgamma;  e3 = log_bads - xdelta;

 /*分散共分散行列の行列式を定義する*/
 det_sig=sigma11*sigma22*sigma33+sigma12*sigma23*sigma13+sigma13*sigma12*sigma23
        -(sigma13**2)*sigma22 - sigma11*(sigma23**2) - (sigma12**2)*sigma33;

 /*分散共分散行列の逆行列の成分(行列式で除していないもの)を定義する*/
 s11 = sigma22*sigma33-sigma23**2; s12 = sigma13*sigma23-sigma12*sigma33;
 s13 = sigma12*sigma23-sigma13*sigma22; s22 = sigma11*sigma33-sigma13**2;
 s23 = sigma12*sigma13-sigma11*sigma23; s33 = sigma11*sigma22-sigma12**2;

 /*尤度関数を定義*/
 ll = -1/2*log(2*pi) +log(1) -1/2*log(det_sig) -1/det_sig
      *(s11*e1**2 + 2*s12*e1*e2 + 2*s13*e1*e3 + s22*e2**2 + 2*s23*e2*e3 + s33*e3**2) ;
 /*尤度関数の最大化を定義*/
 model log_yasai1 ~ general(ll);  run; /*※データセット内にある変数であれば、左辺は何でも良い*/
```

4. 推定結果および考察

3節で述べた推定プログラムを実行することにより、以下の結果を得ることが出来る(表4-1)。

表 4-1　同時推定モデルの推定結果

	保険需要の推定			健康支出の推定			不健康支出の推定		
	推定値	標準誤差	p 値	推定値	標準誤差	p 値	推定値	標準誤差	p 値
保険料				0.2019	0.0328	<.0001	0.2843	0.0666	<.0001
定数項	-0.6320	0.1820	0.0005	6.3060	0.0803	<.0001	6.0533	0.1817	<.0001
大企業	0.0106	0.0169	0.5325	0.1227	0.0071	<.0001	-0.1290	0.0163	<.0001
公務員	0.2529	0.0176	<.0001	0.0769	0.0112	<.0001	-0.1207	0.0240	<.0001
経常所得	0.7417	0.0141	<.0001	-0.0013	0.0256	0.9595	-0.0649	0.0523	0.2144
女性	-0.4343	0.0355	<.0001	0.0352	0.0210	0.0943	-0.6059	0.0455	<.0001
30 代	0.5252	0.0412	<.0001	0.1450	0.0268	<.0001	0.0473	0.0574	0.4096
40 代	0.6862	0.0425	<.0001	0.4166	0.0330	<.0001	0.2188	0.0694	0.0016
50 代	0.7609	0.0424	<.0001	0.6549	0.0357	<.0001	0.3248	0.0747	<.0001
60 代以上	0.6765	0.0463	<.0001	0.9149	0.0338	<.0001	0.2887	0.0712	<.0001
3 人	0.1476	0.0203	<.0001	0.0952	0.0100	<.0001	-0.1531	0.0222	<.0001
4 人	0.2797	0.0207	<.0001	0.1240	0.0132	<.0001	-0.2359	0.0284	<.0001
5 人	0.3318	0.0251	<.0001	0.1826	0.0161	<.0001	-0.1723	0.0346	<.0001
6 人以上	0.4802	0.0377	<.0001	0.2667	0.0235	<.0001	-0.2280	0.0505	<.0001
持家	0.3005	0.0208	<.0001						
ローン完済	-0.1042	0.0273	0.0001						
				N of observation			23759		
				Neg. Log Likelihood			59228		
				AIC			118553		

今回のモデルでは、3 式いずれも被説明変数を自然対数変換している。また、説明変数が連続量のものは自然対数変換を行っていることから、係数値が意味するところは下記の通りとなる。

・連続量の対数値：弾力性(説明変数が 1%変化した際の、被説明変数が X%変化することに相当)

・離散変数：ネイピア数を係数値でべき乗したものが、基準に対する比に相当

以下で推定結果を概観し、解釈を行う。保険料支払い額は、健康支出に対しては想定通りプラスであり、保険料支払が多い世帯ほど健康支出が大きい傾向が確認出来た（保険料 1%の増額に対して、0. 20%の支出増)。一方で不健康支出に関しては、予想と異なりプラスの結果となり、保険料支払が多い世帯ほど不健康支出が

大きい傾向が確認出来た(保険料1%の増額に対して、0.28%の支出増)。この点については、①健康リスクが高い方が却って手厚い保険に加入している可能性、②喫煙者や飲酒者の方が、保険料の支払額が高額に設定されることが影響した可能性、などが考えられる。

その他の説明変数の結果についても、以下に簡単にまとめる。大企業、公務員のいずれも中小企業に対して健康支出は高水準であり、比較的健康への関心が高いことが示唆された。一方で不健康支出は大企業、公務員いずれの世帯でも低水準であり、基本統計量で検討した傾向と不整合な結果となった。また、興味深い点として、経常所得が健康支出、不健康支出のいずれに対しても有意とならなかった。この点については、①生鮮野菜や酒・タバコなどがいずれも所得弾力性が低い可能性、②特にタバコは低所得世帯の方が多く消費するなど、消費額が所得額によってあまり変化しない可能性、③高所得世帯は外食等で消費している可能性、等が考えられる。

この他、女性ダミーについては健康支出に対しては有意な傾向が見られない一方で、不健康支出は有意に少なく、飲酒・喫煙に関する男女の嗜好差が発現したものと考えられる。年代ダミーについては、基本統計量で確認した傾向と整合的に、高齢になるほど健康支出、不健康支出の両方が高まる傾向が見られた。また、世帯人数が増えるにつれ健康支出は増加し、基本統計量と同様の傾向が確認できた。一方で、不健康支出について3人以上の世帯ではいずれも2人世帯よりも支出額が低かった。以上を総合すると健康支出に関しては、基本統計量で確認した傾向と全体的に整合的な結果がモデル推定から得られたと言える。結果の頑健性を確認するために、補足の表4-2～表4-13で2つの世帯属性の組み合わせに関して度数分布と各区分での基本統計量を求めた。健康支出に関しては1属性で見られた傾向が2属性に拡張した場合も大きく崩れる事は無く、強い一貫性が支持されると考えられる。その一方で、不健康支出に関しては強い一貫性があるとは言えず、世帯属性が複雑に関連する事によって推定が不安定になっている可能性が示唆された。この原因として、世帯主の年代が強く関係していると思われる。その理由として、世帯主年齢と他の属性の関係では以下の特徴が確認できる。60歳以上の世帯主は中小企業に勤めている傾向が強く(表4-11)、また世帯人員が少ない傾向が確認できた(表4-12)。この結果から、世帯主が60歳以上の世帯での高い不健康支出水準が強いコホート効果となり、就業状況や世帯人数の影響と混合して推定モデルの結果に影響した可能性が考えられる。

以上のような世帯属性の影響を正確に識別して分析する事は困難である。しかし、疑似ミクロデータの生成元の匿名データでは世帯員の構成も調査されているため、たとえば本稿で定義した不健康支出に影響しない未成年者数を調整することで、精度の高い推定が可能になる可能性がある。

最後に、①式で求めた保険料支払い額の推定結果を見ると、宇野(2013)とほぼ同様の結果が得られた。したがって、全体の1%程度であるゼロ保険世帯を線形モデルで推定したとしても、分析結果に大きく影響しないと考えられる。精緻に分析を行う場合は打ち切りを考慮した分析を行う必要があるが、推定式が複雑になり、推定が収束しない可能性が高まる。

5. まとめと今後の課題

本稿では、世帯のリスク選好度の指標である貯蓄性保険料支払い額と、生鮮野菜・生鮮果物などの健康支出、酒・タバコなどの不健康支出との関係を考察した。同時方程式モデルで推定を行った結果、健康支出に関しては予想と整合的に、保険料支払が多い世帯ほど健康支出が大きい傾向が確認出来た。一方で、不健康支出では予想に反し、保険料支払が多い世帯ほど不健康支出が大きい傾向が確認出来た。その他の世帯属性に関しても、健康支出では基本統計量ベースで検討した傾向と、モデル推定の結果が整合的であった。一方

で、不健康支出については基本統計量で確認した傾向とモデル推定の結果に不一致が数か所見られた。。この原因として、世帯主年齢が不健康支出に対して強いコホート効果として影響している可能性が示唆された。

今回のデータあるいはモデルではこうした可能性の識別は困難であり、今後の課題としたい。今回用いた疑似ミクロデータの生成元である匿名データでは、世帯構成員の詳細データが含まれることから、より詳細な分析が可能と思われる。

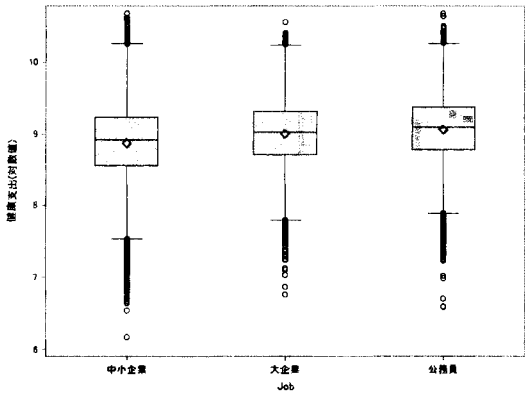
6. 参考・引用文献

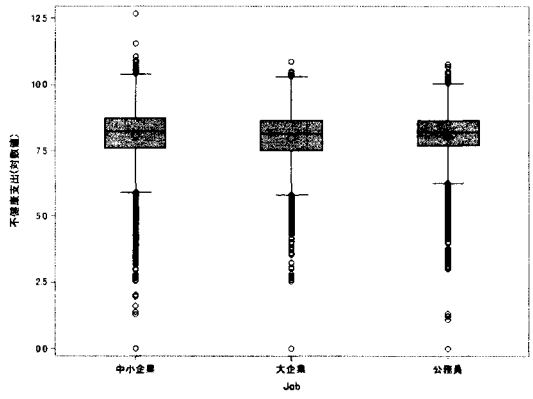
・宇野慧　(2013)　「世帯主の就業状況が貯蓄性保険需要に与える影響についての考察～擬似ミクロデータを用いたTobit/Hurdleモデル推定～」SASユーザー会2013論文集　p.515-518

・Wooldridge, J., (2010), Econometric Analysis of Cross Section and Panel Data (2^{nd} ed.), MIT Press

補足

図 2-2-5　世帯主の就業状況別　健康/不健康支出の箱ひげ図

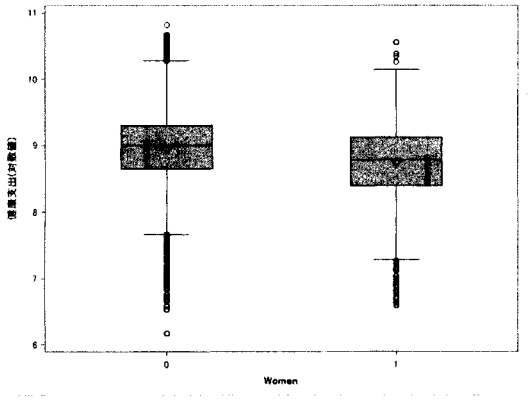


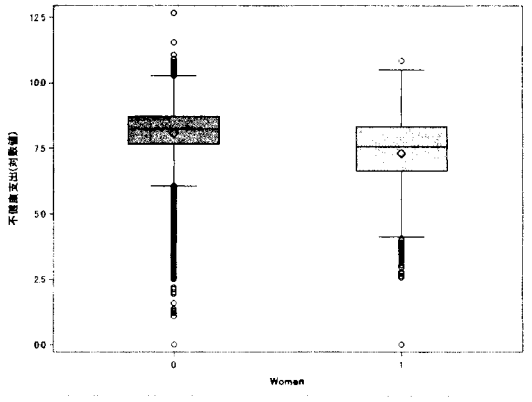


不健康支出	世帯数	mean	sd	median
中小企業	15607	8.88	0.52	8.91
大企業	6999	8.99	0.48	9.02
公務員	6658	9.05	0.48	9.09

不健康支出	世帯数	mean	sd	median
中小企業	15607	8.07	1.08	8.24
大企業	6999	7.99	1.02	8.14
公務員	6658	8.06	0.99	8.21

図 2-2-6　世帯主の性別別　健康/不健康支出の箱ひげ図

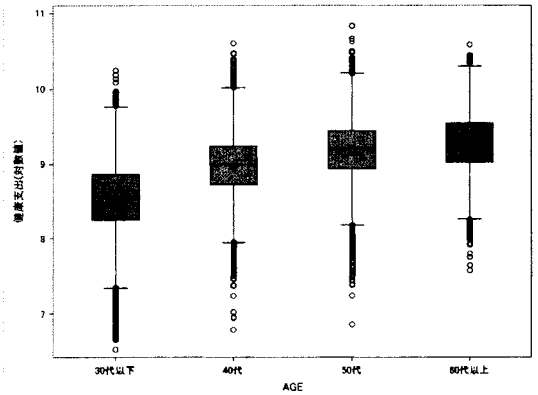


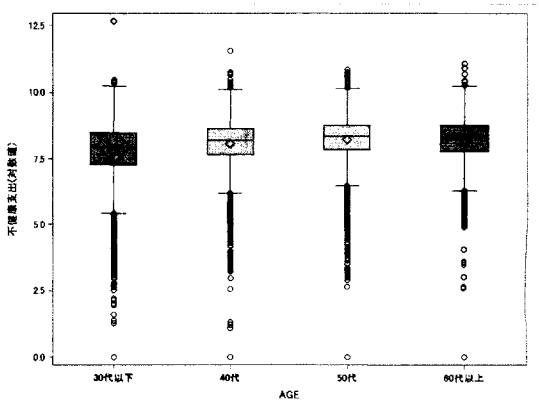


健康支出	世帯数	mean	sd	median
男性	29381	8.95	0.50	8.99
女性	2646	8.74	0.56	8.79

不健康支出	世帯数	mean	sd	median
男性	29381	8.09	1.00	8.23
女性	2646	7.32	1.53	7.58

図 2-2-7　世帯主の年代別　健康/不健康支出の箱ひげ図

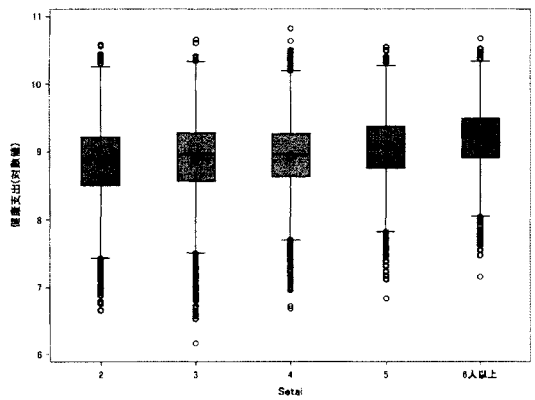


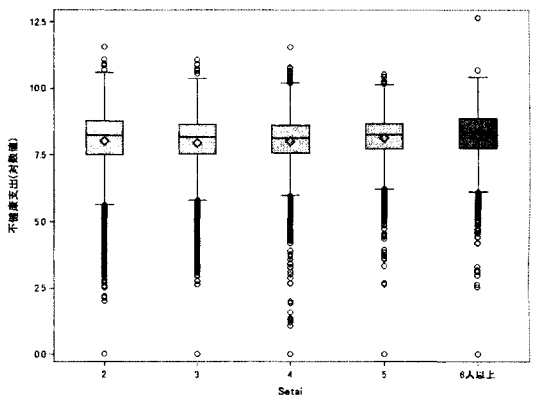


健康支出	世帯数	mean	sd	median
30 代以下	7070	8.54	0.47	8.57
40 代	7831	8.97	0.41	8.99
50 代	8482	9.18	0.40	9.20
60 代以上	2700	9.28	0.39	9.29

不健康支出	世帯数	mean	sd	median
30 代以下	7070	7.79	1.12	7.96
40 代	7831	8.06	1.00	8.19
50 代	8482	8.22	0.96	8.35
60 代以上	2700	8.19	0.97	8.32

図 2-2-8　世帯人員数別　健康/不健康支出の箱ひげ図





健康支出	世帯数	mean	sd	median
2 人	7438	8.84	0.54	8.89
3 人	8537	8.90	0.53	8.96
4 人	9944	8.94	0.48	8.96
5 人	4405	9.05	0.47	9.07
6 人以上	1703	9.20	0.45	9.20

不健康支出	世帯数	mean	sd	median
2 人	7438	8.02	1.20	8.24
3 人	8537	7.96	1.14	8.17
4 人	9944	8.01	1.00	8.14
5 人	4405	8.14	0.89	8.26
6 人以上	1703	8.20	1.02	8.37

表 4-2　世帯主性別および世帯主就業状況に関する度数分布と健康支出の基本統計量表

健康支出		中小企業	大企業	公務員	全体
男性	世帯数(%)	48.87%	23.15%	21.03%	93.05%
	平均値	**8.90**	**9.00**	**9.06**	**8.95**
女性	世帯数(%)	4.46%	0.76%	1.73%	6.95%
	平均値	**8.66**	**8.82**	**8.95**	**8.74**
全体	世帯数(%)	53.33%	23.92%	22.75%	
	平均値	**8.88**	**8.99**	**9.05**	

表 4-3　世帯主性別および世帯主年代に関する度数分布と健康支出の基本統計量表

健康支出		30 代以下	40 代	50 代	60 代以上	全体
男性	世帯数(%)	26.15%	28.32%	30.25%	10.03%	94.74%
	平均値	**8.55**	**8.98**	**9.20**	**9.29**	**8.95**
女性	世帯数(%)	0.95%	1.71%	2.27%	0.32%	5.26%
	平均値	**8.35**	**8.81**	**8.92**	**8.99**	**8.74**
全体	世帯数(%)	27.11%	30.02%	32.52%	10.35%	
	平均値	**8.54**	**8.97**	**9.18**	**9.28**	

表 4-4　世帯主性別および世帯人員数に関する度数分布と健康支出の基本統計量表

健康支出		2 人	3 人	4 人	5 人	6 人以上	全体
男性	世帯数(%)	19.17%	23.99%	29.96%	13.40%	5.21%	91.73%
	平均値	**8.89**	**8.92**	**8.94**	**9.05**	**9.19**	**8.95**
女性	世帯数(%)	4.05%	2.67%	1.09%	0.35%	0.10%	8.26%
	平均値	**8.64**	**8.75**	**8.92**	**9.13**	**9.21**	**8.74**
全体	世帯数(%)	23.22%	26.66%	31.05%	13.75%	5.32%	
	平均値	**8.84**	**8.90**	**8.94**	**9.05**	**9.20**	

表 4-5　世帯主年代および世帯主就業状況に関する度数分布と健康支出の基本統計量表

健康支出		中小企業	大企業	公務員	全体
30 代以下	世帯数(%)	13.77%	7.55%	5.09%	26.41%
	平均値	8.46	8.64	8.65	8.54
40 代	世帯数(%)	14.33%	8.32%	7.62%	30.27%
	平均値	8.88	9.06	9.09	8.97
50 代	世帯数(%)	16.85%	7.90%	8.40%	33.15%
	平均値	9.11	9.26	9.26	9.18
60 代以上	世帯数(%)	7.81%	1.09%	1.26%	10.17%
	平均値	9.25	9.33	9.37	9.28
全体	世帯数(%)	52.76%	24.86%	22.37%	
	平均値	8.88	8.99	9.05	

表 4-6　世帯主年代および世帯人員数に関する度数分布と健康支出の基本統計量表

健康支出		2 人	3 人	4 人	5 人	6 人以上	全体
30 代以下	世帯数(%)	4.24%	8.34%	10.68%	3.19%	0.66%	27.11%
	平均値	8.33	8.48	8.60	8.71	8.96	8.54
40 代	世帯数(%)	2.57%	5.12%	12.89%	6.73%	2.71%	30.02%
	平均値	8.71	8.90	8.97	9.05	9.19	8.97
50 代	世帯数(%)	8.52%	9.96%	8.93%	4.06%	1.06%	32.52%
	平均値	9.02	9.17	9.26	9.33	9.32	9.18
60 代以上	世帯数(%)	5.75%	3.16%	0.95%	0.31%	0.17%	10.35%
	平均値	9.21	9.33	9.39	9.47	9.41	9.28
全体	世帯数(%)	21.09%	26.58%	33.45%	14.29%	4.59%	
	平均値	8.84	8.90	8.94	9.05	9.20	

表 4-7　世帯主就業状況および世帯人員数に関する度数分布と健康支出の基本統計量表

健康支出		2 人	3 人	4 人	5 人	6 人以上	全体
中小企業	世帯数(%)	13.45%	14.64%	15.93%	6.61%	2.69%	53.33%
	平均値	8.85	8.85	8.84	8.97	9.15	8.88
大企業	世帯数(%)	4.56%	6.41%	8.72%	3.50%	0.73%	23.92%
	平均値	8.87	8.96	9.02	9.09	9.22	8.99
公務員	世帯数(%)	5.05%	5.68%	7.03%	3.63%	1.36%	22.75%
	平均値	8.87	9.03	9.07	9.20	9.35	9.05
全体	世帯数(%)	23.07%	26.74%	31.68%	13.73%	4.79%	
	平均値	8.84	8.90	8.94	9.05	9.20	

表 4-8　世帯主性別および世帯主就業状況に関する度数分布と不健康支出の基本統計量表

不健康支出		中小企業	大企業	公務員	全体
男性	世帯数(%)	48.87%	23.15%	21.03%	94.74%
	平均値	**8.14**	**8.00**	**8.12**	**8.09**
女性	世帯数(%)	4.46%	0.76%	1.73%	5.26%
	平均値	**7.21**	**7.54**	**7.37**	**7.32**
全体	世帯数(%)	53.33%	23.92%	22.75%	
	平均値	**8.07**	**7.99**	**8.06**	

表 4-9　世帯主性別および世帯主年代に関する度数分布と不健康支出の基本統計量表

不健康支出		30 代以下	40 代	50 代	60 代以上	全体
男性	世帯数(%)	26.15%	28.32%	30.25%	10.03%	94.74%
	平均値	**7.81**	**8.11**	**8.29**	**8.22**	**8.09**
女性	世帯数(%)	0.95%	1.71%	2.27%	0.32%	5.26%
	平均値	**7.43**	**7.24**	**7.32**	**7.45**	**7.32**
全体	世帯数(%)	27.11%	30.02%	32.52%	10.35%	
	平均値	**7.79**	**8.06**	**8.22**	**8.19**	

表 4-10　世帯主性別および世帯人員数に関する度数分布と不健康支出の基本統計量表

不健康支出		2 人	3 人	4 人	5 人	6 人以上	全体
男性	世帯数(%)	19.17%	23.99%	29.96%	13.40%	5.21%	94.74%
	平均値	**8.19**	**8.03**	**8.03**	**8.15**	**8.20**	**8.09**
女性	世帯数(%)	4.05%	2.67%	1.09%	0.35%	0.10%	5.26%
	平均値	**7.19**	**7.37**	**7.42**	**7.86**	**8.06**	**7.32**
全体	世帯数(%)	23.22%	26.66%	31.05%	13.75%	5.32%	
	平均値	**8.02**	**7.96**	**8.01**	**8.14**	**8.20**	

表 4-11　世帯主年代および世帯主就業状況に関する度数分布と不健康支出の基本統計量表

不健康支出		中小企業	大企業	公務員	全体
30 代以下	世帯数(%)	13.77%	7.55%	5.09%	26.41%
	平均値	**7.85**	**7.70**	**7.78**	**7.79**
40 代	世帯数(%)	14.33%	8.32%	7.62%	30.27%
	平均値	**8.11**	**8.02**	**8.09**	**8.06**
50 代	世帯数(%)	16.85%	7.90%	8.40%	33.15%
	平均値	**8.21**	**8.28**	**8.27**	**8.22**
60 代以上	世帯数(%)	7.81%	1.09%	1.26%	10.17%
	平均値	**8.21**	**8.05**	**8.28**	**8.19**
全体	世帯数(%)	52.76%	24.86%	22.37%	
	平均値	**8.07**	**7.99**	**8.06**	

表 4-12　世帯主年代および世帯人員数に関する度数分布と不健康支出の基本統計量表

不健康支出		2 人	3 人	4 人	5 人	6 人以上	全体
30 代以下	世帯数(%)	4.24%	8.34%	10.68%	3.19%	0.66%	27.11%
	平均値	7.68	7.60	7.89	8.03	8.25	7.79
40 代	世帯数(%)	2.57%	5.12%	12.89%	6.73%	2.71%	30.02%
	平均値	8.19	8.07	7.97	8.13	8.18	8.06
50 代	世帯数(%)	8.52%	9.96%	8.93%	4.06%	1.06%	32.52%
	平均値	8.25	8.22	8.20	8.25	8.20	8.22
60 代以上	世帯数(%)	5.75%	3.16%	0.95%	0.31%	0.17%	10.35%
	平均値	8.19	8.20	8.22	8.10	8.05	8.19
全体	世帯数(%)	21.09%	26.58%	33.45%	14.29%	4.59%	
	平均値	8.02	7.96	8.01	8.14	8.20	

表 4-13　世帯主就業状況および世帯人員数に関する度数分布と不健康支出の基本統計量表

不健康支出		2 人	3 人	4 人	5 人	6 人以上	全体
中小企業	世帯数(%)	13.45%	14.64%	15.93%	6.61%	2.69%	53.33%
	平均値	**8.06**	**8.04**	**8.04**	**8.17**	**8.16**	**8.07**
大企業	世帯数(%)	4.56%	6.41%	8.72%	3.50%	0.73%	23.92%
	平均値	**8.07**	**7.99**	**7.91**	**8.01**	**8.30**	**7.99**
公務員	世帯数(%)	5.05%	5.68%	7.03%	3.63%	1.36%	22.75%
	平均値	**8.06**	**7.92**	**8.07**	**8.21**	**8.20**	**8.06**
全体	世帯数(%)	23.07%	26.74%	31.68%	13.73%	4.79%	
	平均値	**8.02**	**7.96**	**8.01**	**8.14**	**8.20**	

第2回ミクロデータ分析コンテスト 規定課題

木下 陽介，若林 将史，飯塚 政人，中川 雄貴
東京理科大学大学院 工学研究科 経営工学専攻
カテゴリー B

The compulsory program in the contest

Yosuke Kinoshita, Masashi Wakabayashi, Masato Iizuka, Yuki Nakagawa
Department of Management Science, Graduate School of Engineering, Tokyo University of Science

1. 第 1-1 表

作成したプログラムと出力結果は以下の通り．

```
proc freq data = data noprint;
  tables ShuugyouJinin*SetaiJinin / out = out1(drop =
percent);
  where ShuugyouJinin in ( 1, 2) and SetaiJinin in (4);
run;
proc print data = out1 label; run;
```

OBS	有業人員	世帯人員	度数
1	1	4	4132
2	2	4	4201

2. 第 1-2 表

作成したプログラムと出力結果は以下の通り．

```
proc freq data = data noprint;
   tables ShuugyouJinin*SetaiJinin / out = out2(drop =
   percent);
   where ShuugyouJinin in ( 1, 2) and SetaiJinin in (4);
   weight Weight; run;
data out2; set out2;
  r_count = round( count, 1.);
  label r_count = "修正度数"; run;
proc print data = out2 label; run;
```

OBS	有業人員	世帯人員	度数	修正度数
1	1	4	66299.27	66299
2	2	4	64868.23	64868

3. 第 1-3 表

作成したプログラムと出力結果は以下の通り．

```
proc freq data = data noprint;
  tables ShuugyouJinin*SetaiJinin / out = out3(drop =
  count);
  weight Weight; run;
data out3; set out3;
  count10 = percent * 1000;
  r_count10 = round( count10, 1.);
  label count10 = "度数" r_count10 = "修正度数";
  drop percent; run;
proc print data = out3 label;
  where ShuugyouJinin in ( 1, 2) and SetaiJinin in
  (4);run;
```

OBS	有業人員	世帯人員	度数	修正度数
3	1	4	13381.23	13381
10	2	4	13092.40	13092

4. 第 2 表

作成したプログラムと出力結果は以下の通り．ただし，出力結果の一部は加工を施している．

```
%macro means_mcr( data_out, where);
proc freq data = data noprint;
  tables SetaiJinin / out = temp01(drop = percent);
  where &where.;
  weight Weight; run;
data temp01; set temp01;
```

```
  r_count = round( count, 1);
  label r_count = "世帯数分布（抽出率調整）"; run;
proc means data = data noprint;
  where &where.;
  var Syouhishisyutu Food House Kounetsu_Suidou
  Kagu_Kaji Hihuku Hoken_Iryou Koutsuu_Tsuushin
  Kyouiku Kyouyou Other_Syouhishisyutsu;
  weight Weight;
  output out = temp11(drop = _TYPE_ _FREQ_); run;
data temp12; set temp11;
  where _STAT_ in ("MEAN");
  array youto{11} Syouhishisyutu Food House
  Kounetsu_Suidou Kagu_Kaji Hihuku Hoken_Iryou
  Koutsuu_Tsuushin Kyouiku Kyouyou
  Other_Syouhishisyutsu;
  array r_youto{11} r_Syouhishisyutu r_Food r_House
  r_Kounetsu_Suidou r_Kagu_Kaji r_Hihuku
  r_Hoken_Iryou r_Koutsuu_Tsuushin r_Kyouiku
  r_Kyouyou r_Other_Syouhishisyutsu;
  do i = 1 to 11;
    r_youto{i} = round( youto{i}, 1.);
  end;
  label r_Syouhishisyutu = "消費支出" r_Food = "食料"
  r_House = "住居" r_Kounetsu_Suidou = "光熱・水道"
  r_Kagu_Kaji = "家具・家事用品" r_Hihuku = "被服及
  び履物" r_Hoken_Iryou = "保険医療"
  r_Koutsuu_Tsuushin = "交通・通信" r_Kyouiku = "教
  育" r_Kyouyou = "教養娯楽"
  r_Other_Syouhishisyutsu = "その他の消費支出";
  drop _STAT_ Syouhishisyutu Food House
  Kounetsu_Suidou Kagu_Kaji Hihuku Hoken_Iryou
  Koutsuu_Tsuushin Kyouiku Kyouyou
  Other_Syouhishisyutsu i; run;
data temp21; set temp11;
  where _STAT_ in ("N");
  keep where Syouhishisyutu; run;
data temp22; set temp21;
  rename Syouhishisyutu = n;
  label n = "集計世帯数"; run;
data &data_out.;
  length where $50.;
  where = "&where."; run;
data &data_out.; merge &data_out. temp22 temp01(keep
= r_count) temp12; run;
%mend means_mcr;
%means_mcr( out001, ShuugyouJinin in (1) and
SetaiJinin in (4));
%means_mcr( out002, ShuugyouJinin in (2) and
SetaiJinin in (4));
%means_mcr( out003, ShuugyouJinin in ( 1, 2) and
SetaiJinin in (4));
data out010; set out001-out003; run;
proc print data = out010 label; run;
```

抽出条件*	消費支出	世帯数分布（抽出率調整）	消費支出	食料	住居
1	4132	66299	305234	71543	17556
2	4201	64868	347740	78472	12932
3	8333	131168	326256	74970	15269

抽出条件*	光熱・水道	家具・家事用品	被服及び履物	保険医療	交通・通信
1	18854	8383	13579	11656	42703
2	20621	8917	14876	10538	52141
3	19728	8647	14221	11103	47371

抽出条件*	教育	教養娯楽	その他の消費支出
1	31202	31959	57801
2	39485	33681	76078
3	35298	32810	66839

* 1：就業人員 1人，世帯人員 4人
2：就業人員 2人，世帯人員 4人
3：就業人員 1,2人，世帯人員 4人

ジニ係数による所得格差の解析

木下 陽介，若林 将史，飯塚 政人，中川 雄貴
東京理科大学大学院 工学研究科 経営工学専攻

カテゴリーB

The analysis of income gap by Gini coefficient

Yosuke Kinoshita, Masashi Wakabayashi, Masato Iizuka, Yuki Nakagawa
Department of Management Science, Graduate School of Engineering, Tokyo University of Science

要旨

近年の就職活動の傾向として，学生の大手志向が注目されている．また，公務員もその安定性から就職先として人気のある業種である．本研究では，多角的な視点から，勤め先収入の中央値とジニ係数を算出し，企業規模の大きさや産業が勤め先収入とその格差に与える影響を解析した．本稿では，企業規模の他に，産業，性別，年齢などが，勤め先収入の高さとその所得格差へ与える影響についての結果，考察を示す．その結果，企業規模，産業によって所得格差に影響を与えることがわかった．また，医療・福祉は特に所得格差が大きく，この原因について解決するため，更なる追加解析をおこなった．

キーワード：所得格差，ジニ係数，年齢，労働環境，企業区分，企業規模，性別，SGPLOT

1. 背景

今日，学生の企業選択基準のひとつに「大手企業」という指標がある．大手企業が好まれる理由として，「働きがいのある仕事がしたい」，「責任のある仕事がしたい」といった理由に加え，「安定して高所得を得たい」という意見が多く挙げられる．しかし，実際に「大手企業であれば安定した高収入が見込めるか」ということには検討の余地がある．

我々は，この問題について調べるため，企業規模を始めとし，様々な方面から，給与所得について調べ，所得格差についてみるため，ジニ係数を算出した．ジニ係数は格差の指標として用いられる．集中・独占・不平等といった傾向を見出す指標として，特に所得分布を表す場合に広く利用されている．本研究では，このジニ係数を用い，所得格差について調査をおこなった．

2. 目的

所得格差に影響しうる要因として，企業規模に加え，産業，性別，年齢等が考えられる．そこで，各要因別に勤め先収入の中央値，ジニ係数を算出し，給与水準，及び，所得格差の違いについて調べ，比較することを目的とした．今回は，勤労者が１人(世帯主 = 勤労者)である世帯に絞り，給与所得として勤め先収入を用いた解析をおこなった．

3. 方法

3.1. 解析対象

本研究では勤め先収入を対象とし，勤労者が 1 人である世帯に絞り解析をおこなった．勤労者が 1 人である世帯に絞った理由として，勤め先収入は世帯ごとの合計値で表されるため，勤労者の人数が異なると，見かけ上所得格差が大きく算出されると考えたためである．また勤め先収入の分布は非常に偏っているため，比較のために，外れ値の影響を受けづらい中央値を算出した．

本研究では，解析の対象となる因子として，企業規模，産業，性別，年齢を用いた．

3.2. ジニ係数

ジニ係数は，所得格差を見る際に用いられる代表的な指標の 1 つである．ここでジニ係数と関連の深い，ローレンツ曲線について述べ，ジニ係数の算出方法について説明する．

3.2.1. ローレンツ曲線

ローレンツ曲線は所得の不平等や集中の度合いを観察するために，度数分布表から作られる曲線である．ローレンツ曲線は，横軸に低所得世帯からの累積相対度数，縦軸に所得の累積相対度数を目盛り，対応する点を順次結ぶことで描くことができる．したがってローレンツ曲線は 0 から 1 まで単調に増加する曲線となる．もしすべての世帯が同一の収入であれば，ローレンツ曲線は 45 度線と一致する．この 45 度線のことを完全平等線という．

3.2.2. ジニ係数の算出方法

ジニ係数は，ローレンツ曲線と完全平等線で囲まれた面積の 2 倍と定義される．本研究では度数分布表からローレンツ曲線の下側の面積を台形の面積の総和として求め，ジニ係数を算出した．尚，平成 16 年度の全国消費実態調査結果から，勤労者世帯のジニ係数は 0.257 であった．我々が算出したジニ係数は 0.243 であり，擬似ミクロデータからもジニ係数がある程度再現できると判断した．

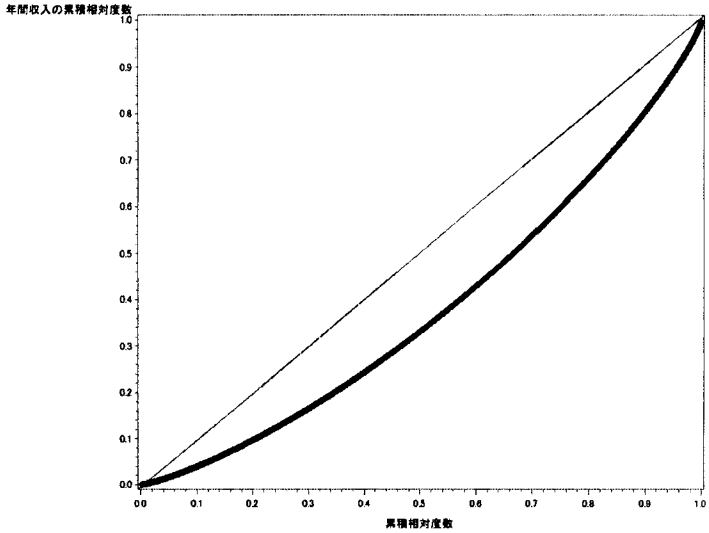


図 1．勤労者世帯における年間収入のローレンツ曲線

4. 解析結果・考察

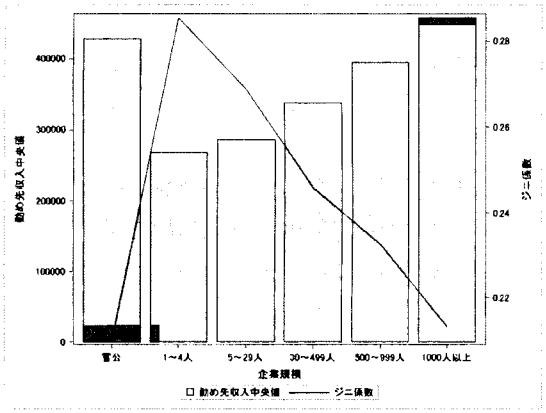


図 2. 企業規模別の勤め先収入中央値とジニ係数

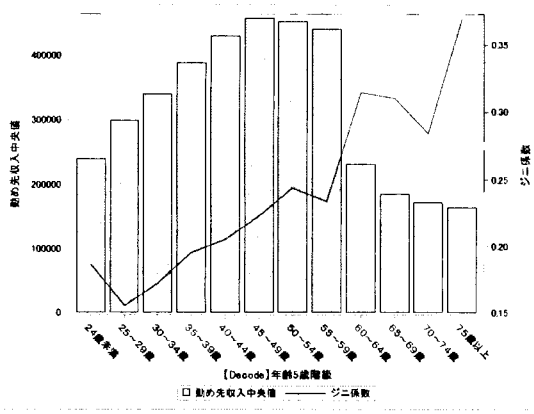


図 3. 年齢 5 歳階級の勤め先収入中央値とジニ係数

表 1. 産業区分の解析結果

産業	勤め先収入の中央値(円)	ジニ係数	N	男性比率
飲食店・宿泊業	133548	0.263	684	36.4
不動産業	234879	0.276	510	100.0
医療・福祉	302243	0.357	4920	48.6
建設業	328991	0.241	12496	98.1
サービス業（他に分類されないもの）	342899	0.263	13968	93.4
卸売・小売業	345234	0.256	16661	89.2
運輸業	345968	0.219	7231	100.0
複合サービス事業	366410	0.195	1298	100.0
製造業	389719	0.221	39577	96.0
公務（他に分類されないもの）	428968	0.194	23557	94.7
教育・学習支援業	463112	0.212	7312	87.6
情報通信業	466461	0.188	6621	100.0
電気・ガス・熱供給・水道業	477757	0.223	1652	100.0
金融・保険業	489375	0.266	4393	91.6

※勤め先収入の中央値の高い順

表 2. 性別区分の解析結果

性別	勤め先収入の中央値(円)	ジニ係数	N
女	166338	0.360	22146
男	384194	0.233	197596

※勤め先収入の中央値の高い順

図 2 で企業の規模別に勤め先収入の中央値とジニ係数を算出したところ，企業規模が大きくなるほど収入が高まり，ジニ係数が低くなることが確認できた．最も収入が高く，ジニ係数が少ないのは，官公であった．したがって，最近の学生に多く見られる「安定・高収入」を求めての大手志向，あるいは公務員の人気は妥当なものであると判断できる．

図 3 を見ると， 24 歳未満から 55~59 歳に至るまで，勤め先収入・ジニ係数ともに上昇傾向にあることがわかる．勤め先収入に上昇傾向がみられた要因として，今日の日本において年功序列の慣習の影響が大きいと考えられる．また，ジニ係数が上昇傾向にある理由に，年齢に対する給与の推移が挙げられる．入社して間もないうちは，多くの者は変わらない収入を得ているのに対して，歳を取るにつれ，その人の能力や出世の度合いにより収入の格差が開いていくのではないかと予想できる．60 歳以降，急激に収入が減る理由として，定年の影響が考えられる．さらに 60 歳以降，ジニ係数が急激に高まっていることから，収入の格差がより開いていることがわかる．このようになった原因のひとつとして，60 歳以上の就業者に対して，定年後に再就職をして収入が減ったものと，60 歳以前とそう変わらない収入を得ているものがいる影響で，ジニ係数が高まった可能性が考えられる．平成 26 年度現在，定年は 65 歳に引き上げられつつあり，現在の年齢による勤め先収入への影響は興味深い．

表 1 は産業別に勤め先収入の中央値とジニ係数を算出したものである．これより，医療・福祉，情報通信業，公務で特徴的な結果が得られた．まず，医療・福祉では他の産業に比べ，ジニ係数が非常に高いことがわかった．情報通信業では勤め先収入が高く，ジニ係数が他の産業に比べて最も低い値であった．同じく，公務に当たるものも勤め先収入が高く，ジニ係数が低いという傾向がみられた．公務については前述した企業規模別の官公のものと同様の結果が得られており，やはり安定・高収入の面で優れていると考えられる．

情報通信業は月収が高く，ジニ係数が低いことから安定・高収入の面で優れている可能性がある．しかし，情報通信業は男性比率が 100.0%となっており，女性の勤労者が存在していない．性別に勤め先収入を比べた時，男性は女性に比べ収入が多い傾向にあり，情報通信業では男性のみのデータを取り扱うことから，ジニ係数が減り格差が減ったようにみられた可能性がある．

医療・福祉ではジニ係数が最も高い結果が得られた．このような結果が得られた理由の一つに医療・福祉では男性比率が 48.6%と他の産業に比べ少ないことが原因だと考えられる．この産業で勤労者の女性比率が高い理由として，女性看護師が多いことが考えられる．医師と看護師では収入に差が大きく，ジニ係数が高くなった可能性がある．そのため，医療・福祉について給料格差が大きい原因を探るため，性別に給与の分布について，追加解析をおこなった．

5. 追加解析

表 3. 医療・福祉の性別にみた解析結果

性別	勤め先収入の中央値(円)	ジニ係数	N
女	223938	0.297	2532
男	425918	0.319	2389

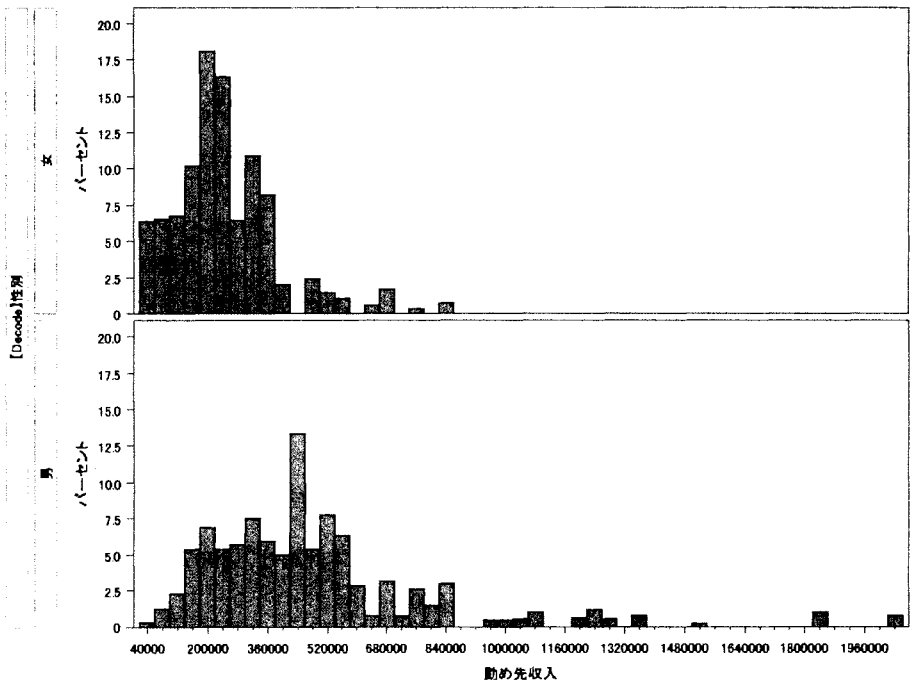


図 4. 医療・福祉の性別ごとにみた勤め先収入の中央値のヒストグラム

医療・福祉で，性別に勤め先収入の中央値を算出したところ，女性が 22 万円程度であるのに対して，男性は 42 万円と大きな差がみられた．前述したように男女比率も均等に近いため，ジニ係数が高まったものと考えられる．また，性別ごとにジニ係数を算出すると男性が高い傾向にある．これは他の産業におけるジニ係数の傾向とは異なる．図 4 は医療・福祉の性別ごとにみた勤め先収入の中央値についてのヒストグラムである．これにより，男性でより広く分布していることが確認できた．このようになった理由として，医療・福祉の中で収入の高い集団と低い集団が混合していることが挙げられる．医療・福祉に従事する男性は，医師などの高収入な集団と，介護職などの比較的収入が低い集団が混合していることが予想できる．したがって，男性の中での収入格差が高い傾向にあったのではないかと推察する．このことも医療・福祉全体で見たジニ係数の高さに影響をしていると考える．

6. まとめ

本研究では，多角的な視点から，勤め先収入の中央値とジニ係数を算出し，企業規模の大きさや産業が勤め先収入とその格差に与える影響を調査した．所得格差に影響しうる要因として，企業規模に加え，産業，性別，年齢等が考えられる．そこで，各要因別に給与の中央値，ジニ係数を算出し，給与水準，及び，所得格差の違いについて調べ，比較した．ただし，ジニ係数は不平等を表す指標ではあるが，調査対象に特定の傾向がある場合，不平等が必ずしも悪いことになるとは限らない点に留意する必要がある．

年齢が上昇するにつれ，勤め先収入の中央値に増加傾向がみられた．これは，年功序列の慣習が色濃く残っていると考えられる．同様に，ジニ係数にも上昇傾向がみられた．年齢を重ねるごとに，その人の能力や出世の度合いにより収入の格差が開いていくためにこのような結果となったのだと推察できる．企業規模が大きくなるほど収入が高まり，ジニ係数が低くなることが確認できた．最も収入が高く，ジニ係数が少ないのは官公であった．したがって，最近の学生に多く見られる「安定・高収入」を求めての大手志向，あるいは公務員志向は妥当なものであると判断できた．産業別にみると，医療・福祉では他の産業と比較して，ジニ係数が非常に高いことがわかった．この原因として，男性比率が他の産業に比べ少ないことが考えられる．追加解析の結果，医療・福祉に従事する男性は，医師などの高収入な集団と，介護職などの比較的収入が低い集団が混合していることが推察された．したがって，医療・福祉のジニ係数を高めたと結論付ける．

7. 参考文献

[1] 作間 逸雄．SNA がわかる 経済統計学．有斐閣アルマ．2003.

[2] 白砂堤津耶．例題で学ぶ 初歩からの計量経済学 第 2 版．日本評論社．2007.

[3] 田中 勝人．現代経済学入門 経済統計．岩波書店．2009.

[4] 平成 16 年度全国消費実態調査「http://www.stat.go.jp/data/zensho/2004/」
(最終閲覧日:2014 年 5 月 27 日)

[5] マイナビ採用サポネット | 新卒企画「http://saponet.mynavi.jp/enq_gakusei/ishiki/index.html」
(最終閲覧日:2014 年 5 月 27 日)

擬似ミクロデータにおける集計表の再現

芥川　麻衣子　（参加カテゴリー：C）
株式会社タクミインフォメーションテクノロジー
システム開発推進部　数理解析グループ

1. データ読み込み

INFILE ステートメントで読み込みを行った。一連の DATA ステップには、マクロプログラムを適用した。読み込んだ 7 つのデータセットを縦結合して 1 つのデータセットにした。

2. 集計表の再現

2-1　第 1-1 表　集計世帯数（各レコードを、単純にカウントしたもの）

FREQ プロシジャを用いた。結果は表 1 の通りである。

2-2　第 1-2 表 世帯数分布（各レコードを、集計用乗率で重み付けして、カウントしたもの。）

FREQ プロシジャに、WEIGHT ステートメントで重みづけをした。データセットに出力し、TRANSPOSE プロシジャ等で成形した。結果は表 2 の通りである。

2-3　第 1-3 表 世帯数分布（10 万分比）

表 2 の値を 495,465 で割り、100,000 を掛けた。結果は表 3 の通りである。

2-4　第 2 表　支出（消費支出 及び 十大費目）

まず、有業人員 1 又は 2 人のフラグ、重みづけした世帯と支出をデータセットに作成。次に、総数、世帯人員 4 人、有業人員 1 又は 2 人に分けて、支出の平均を算出。平均の算出は、MEANS プロシジャで集計の後、DATA ステップ内でループ処理にて行った。最後に、3 つのデータセットを縦結合し、表の形を得た。結果は表 4 の通りである。

表 1　集計世帯数【再現】

表：ShuugyouJinin * SetaiJinin										
ShuugyouJinin(有業人員)	SetaiJinin(世帯人員)									
度数	2人	3人	4人	5人	6人	7人	8人	9人	10人	合計
1人	4124	3908	4132	1436	256	51	6	0	0	13913
2人	3239	3391	4201	1943	494	162	29	0	0	13459
3人	0	1035	1031	559	232	84	6	3	0	2950
4人	0	0	324	220	104	31	12	0	0	691
5人	0	0	0	27	6	7	0	0	0	40
6人	0	0	0	0	3	3	0	0	0	6
不詳	75	203	256	220	119	52	28	12	3	968
合計	7438	8537	9944	4405	1214	390	81	15	3	32027

表 2　分布世帯数【再現】

有業人員	2人	3人	4人	5人	6人	7人	8人	9人	10人
1人	64,691	61,284	66,299	22,740	3,813	831	84	0	0
2人	50,138	51,523	64,868	29,616	6,801	2,336	441	0	0
3人	0	15,912	15,615	7,963	3,302	1,137	76	36	0
4人	0	0	4,851	3,345	1,354	422	195	0	0
5人	0	0	0	383	80	108	0	0	0
6人	0	0	0	0	49	51	0	0	0
不詳	1,006	3,287	4,216	3,519	1,761	727	401	170	33

表 3　世帯数分布（10 万分比）【再現】

有業人員	2人	3人	4人	5人	6人	7人	8人	9人	10人
1人	13,057	12,369	13,381	4,590	770	168	17	0	0
2人	10,119	10,399	13,092	5,977	1,373	471	89	0	0
3人	0	3,211	3,152	1,607	666	229	15	7	0
4人	0	0	979	675	273	85	39	0	0
5人	0	0	0	77	16	22	0	0	0
6人	0	0	0	0	10	10	0	0	0
不詳	203	664	851	710	355	147	81	34	7

表 4　支出（消費支出 及び 十大費目）【再現】

Shuugyoujinn	集計世帯数	世帯数分布(消費支出)	消費支出	食料	住居	光熱・水道
総数	32,027	495,465	328,140	72,883	17,687	19,238
うち世帯人員が4人	9,944	155,850	335,438	76,362	15,345	20,214
有業人員1人	4,132	66,299	305,234	71,543	17,556	18,854
有業人員2人	4,201	64,868	347,740	78,472	12,932	20,621
有業人員3人	1,031	15,615	380,521	83,796	12,583	23,045
有業人員4人	324	4,851	399,962	85,083	18,500	22,490
不詳	256	4,216	379,882	82,134	24,296	22,238
(特掲)有業人員1人又は2人	8,333	131,168	326,256	74,970	15,269	19,728

（表４続き）

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
9,204	14,138	11,366	47,961	22,270	31,389	82,003
8,885	14,452	10,987	47,894	33,442	32,269	75,588
8,383	13,579	11,656	42,703	31,202	31,959	57,801
8,917	14,876	10,538	52,141	39,485	33,681	76,078
10,276	16,561	10,331	52,406	22,513	27,852	121,160
11,763	14,096	10,812	51,310	4,509	32,769	148,628
7,843	14,238	10,011	43,521	49,453	31,206	94,942
8,647	14,221	11,103	47,371	35,298	32,810	66,839

ミクロデータを用いたセルフメディケーション の実態把握に関する検討

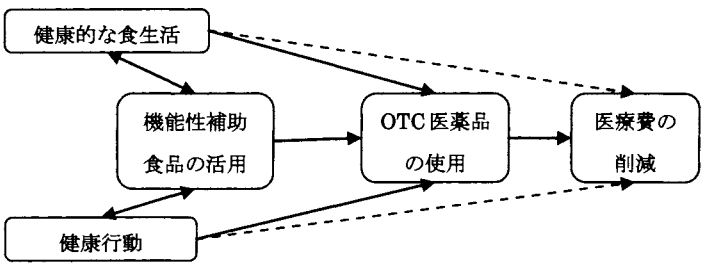
芥川　麻衣子　（参加カテゴリー：C）
株式会社タクミインフォメーションテクノロジー
システム開発推進部　数理解析グループ

1. はじめに

セルフメディケーションとは、「自分自身の健康に責任を持ち、軽度な身体の不調は自分で手当てすること」と WHO は定義している。このような観点から、OTC（Over The Counter）医薬品の活用と合わせて語られることが多い[1]。そして、セルフメディケーションによって、医療費の削減が期待される。しかし、市民が自身の健康を管理した上で、OTC 医薬品使用の適正な判断ができているかについては、議論の余地がある。そこで本研究は、ミクロデータを用い、各世帯における生活習慣、OTC 医薬品の使用、医療費の関係について実態把握することを目的とする。

2. 研究の概念

研究の概念を図 1 に示す。前項で述べた通り、セルフメディケーションは、広義には自分の健康に責任を持った生活習慣、狭義には OTC 医薬品の活用である。健康的な生活をした上で、OTC 医薬品が利用されることが望ましい。また、セルフメディケーションは、医療費の削減に繋がる。そこで、本研究では生活習慣として、ミクロデータより取得可能な情報から、「健康的な食生活」「機能性補助食品の活用」「健康行動」を項目とする。これらと、OTC 医薬品の活用との関係について確認する。アウトカムとして、医療費との関係を確認する。



（本研究の対象となる関係を実線、対象外を破線で示している。本稿は、生活習慣による医療費への直接影響ではなく、OTC 医薬品の使用を介した間接影響に着目する。）

図 1　研究の概念図

3. 方法

3-1　変数の採用

研究の概念に基づき、使用する変数について説明する。健康的な食生活として、食料の項目を使用する。健康的な食生活をしている世帯は、各食材をバランスよく購入していることが予想される。そこで、食料に関する項目を主成分分析し、第一主成分をその指標とした。機能性補助食品の活用として、健康保持用摂取品の項目を使用する。これは、栄養成分の補給など保健、健康増進のために用いる食品と定

義されており、サプリメントなどがこれに類する[2)]。健康行動として、たばこの項目を使用する。OTC医薬品の活用として、医薬品の項目を使用する。但し、当該項目には、医師の処方箋により院外で購入した薬も含まれている[2)1]。医療費として、保健医療サービスの項目を使用する。こちらについても、世帯が負担した額がベースであり[2)]、負担率の影響を受けている点に留意が必要である。属性として、年間収入を用いた。健康と経済社会的要因との関係が先行研究により示されているためである[2]。

3-2 属性

各世帯の属性として、年間収入を 7 階級に分けた。階級範囲には、四分位点に加えて、低所得者層は下位 10%を分離し、高所得層については大きな幅があったため、上位 10%と 2,000 万円以上を分けた[3]。各階級における年間収入の要約統計量は表 1 の通りである。

表 1 収入階級の要約統計量

(千円) (千円)

階級	範囲	分位点	N	平均	標準偏差	中央値
最高収入階級	20000以上	-	157	24976.52	7941.28	22274.95
高収入階級	11491.597以上	90%	2882	13363.07	1509.05	12946.76
高～中収入階級	8989.082以上	75%	4804	10070.44	712.57	9983.85
中収入階級	6719.73以上	50%	8007	7767.52	648.30	7724.90
中～低収入階級	4959.886以上	25%	8007	5833.63	507.57	5830.52
低収入階級	3686.886以上	10%	4804	4365.82	364.14	4390.15
最低収入階級	3686.886より低い	-	3202	2819.85	696.01	2990.35

3-3 食生活指標

主成分分析には、飲料を除く食料の下位項目を用いた[4]。結果、第一主成分の寄与率は 44.3%、固有ベクトルは表 2 の通り全て正の値であった。よって、第一主成分は食材を満遍なく摂取している指標に代えることができる。さらに主成分得点を算出した。各階級の第一主成分得点の平均値は表 3 の通りである。正の値ならば満遍なく摂取できており、負の値ならばできていないと解釈できる。また、健康的な食生活の変数として、これを使用している。

表 2 第一主成分の固有ベクトル

項目	値	(続き) 項目	値
米	0.16	卵	0.23
パン	0.20	生鮮野菜	0.25
めん類	0.21	乾物・海藻	0.20
他の穀類	0.15	大豆加工品	0.24
生鮮魚介	0.23	他の野菜・海藻加工品	0.22
塩干魚介	0.21	果物	0.20
魚肉練製品	0.22	果物加工品	0.10
他の魚介加工品	0.20	油脂	0.18
生鮮肉	0.24	調味料	0.27
加工肉	0.22	菓子類	0.21
牛乳	0.19	主食的調理食品	0.16
乳製品	0.13	他の調理食品	0.19

表 3 収入階級の第一主成分得点の平均値

最高収入階級	高収入階級	高～中収入階級	中収入階級	中～低収入階級	低収入階級	最低収入階級
1.26	0.75	0.48	0.17	-0.19	-0.47	-0.74

[1]平成 24 年度の市場規模は、医療用医薬品 8 兆 7,660 億円に対し、一般用医薬品 7,161 億円である。(厚生労働省 薬事工業生産動態統計調査より。市場規模は、国内出荷額を使用している。尚、同調査における一般医薬品とは、配置用家庭薬を除く。) 国民医療費の薬剤比率は、20%台である。(厚生労働省 医薬品産業ビジョン 2013 より) 保健医療調剤における院外処方の割合は、日本薬剤師会の処方せん受取率で知ることができるが、平成 24 年度時点、全国で 65.1%である。平成 21 年に厚生労働省が算出した院外処方率によると、全体で 62.0%、病院は 70.0%、診療所は 59.0%とある。

[2]社会経済的属性として、収入と就労状況が考えられる。平成 19 年の内閣府 経済社会総合研究所による「健康と経済社会的属性との関係に関する調査研究報告書」もこれらの関係を、個票データより検討している。報告によると、個人内では、収入の増減よりも就労状況の変化が健康度の変化と関係しており、これは健康度が下がった人が非就労へ移行していることによると考察されている。一方、集団間では、収入水準と健康度が相関していた。よって、本研究は一時点の個票データであるため、収入を属性とした。

[3]99%分位点(17681.528 千円)を検討したが、1000 万円台は数が多く、2000 万円から急激に減り、7000 万円まで分布していたため、2000 万円以降として分離することとした。

[4]擬似ミクロデータでは明示されていないが、食料項目は中分類と小分類に分かれる (例えば、中分類穀類の中に、米、パンなどが入っている)。中分類項目は小分類と 0.8 以上の相関があったことから、除外した。

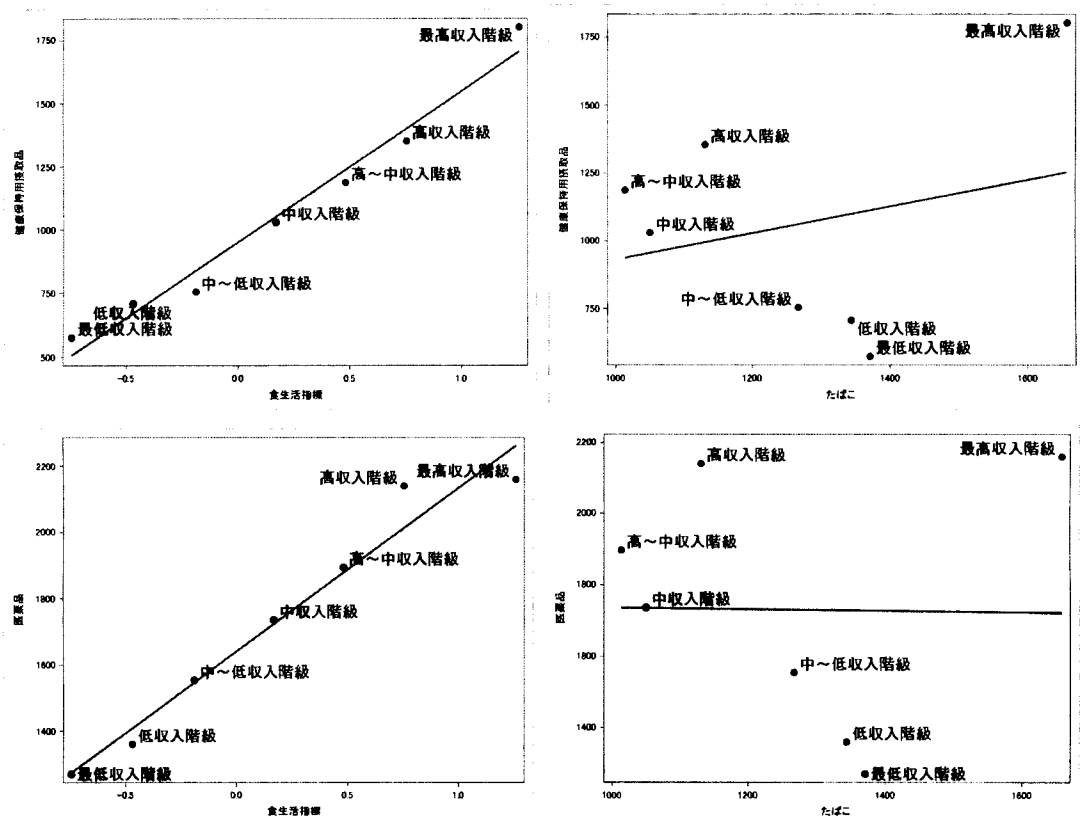
3-4　各項目の関連

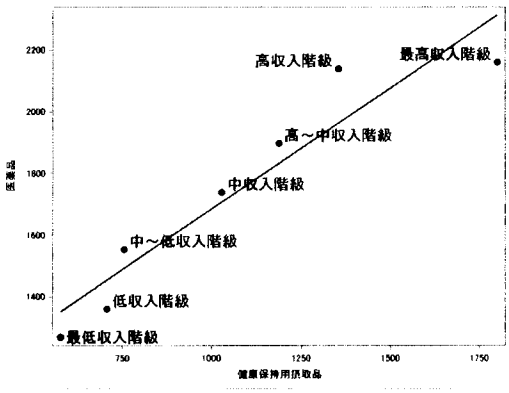
3-1 で定義した変数同士の関連を、散布図により確認する。まず、集団傾向を把握するため、収入階級別に支出額及び主成分得点の平均値をプロットし、回帰直線をひく。回帰直線は傾向線として、プロットが直線より上にあるものは x 軸項目への支出傾向に比べ y 軸項目への支出が多く、プロットが直線より下にあるものは x 軸項目への支出傾向に比べ y 軸項目への支出が少ないと読むことができる。

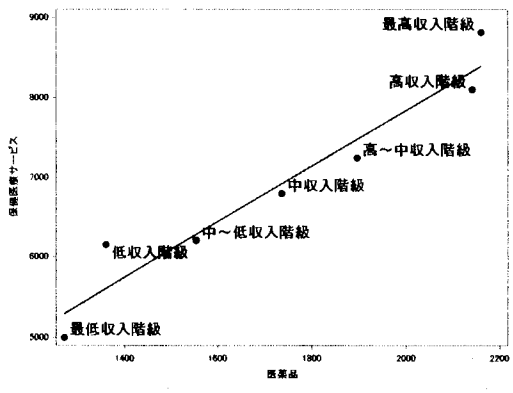
次に、中所得層の 3 階級について、階級内の個票をプロットする。散布図の外れ値を検討し、セルフメディケーションの観点から考察する。尚、散布図を作成するにあたり、食生活指標以外の支出項目は対数変換を行った。そのため、支出が 0 の個票は除外されている。食生活指標である主成分得点は基準化して正規分布に近い形であったため、対数変換していない。

4. 結果と考察

4-1　収入階級の散布図







4-2　収入階級別の考察

たばこ以外の項目同士で正の相関が見られた[5]。また、各項目への支出額が多いほど、収入が高いという関係も見られた。よって、健康的な食生活の上に機能性補助食品を活用し、さらに OTC 医薬品が活用されていることが示唆される。

ここからは、項目間について考察する。健康的な食生活と機能性補助食品の活用では、低収入層は食生活が良好でなく、機能性補助食品摂取に偏っており、コンビニ食に機能性補助食品をプラスする食生活が想像される。中収入層以上は機能性補助食品に偏らない傾向が見られたが、最高収入階級で機能性補助食品の摂取が高かった。

たばこは機能性補助食品の摂取、OTC 医薬品の使用に関連がなかった。しかし、最高収入階級が相関関係に大きな影響を与えていることが散布図から解釈できる。そのため、最高収入階級を除くと負の相関があることが予想される。つまり、喫煙をしない、健康行動をしている人ほど、機能性補助食品や OTC 医薬品を活用している。この点について、仮説を支持する。また、たばこは低所得者層ほど支出額が多い傾向も見られた。これは、平成 22 年度の国民健康・栄養調査の結果と一致する [3)]。

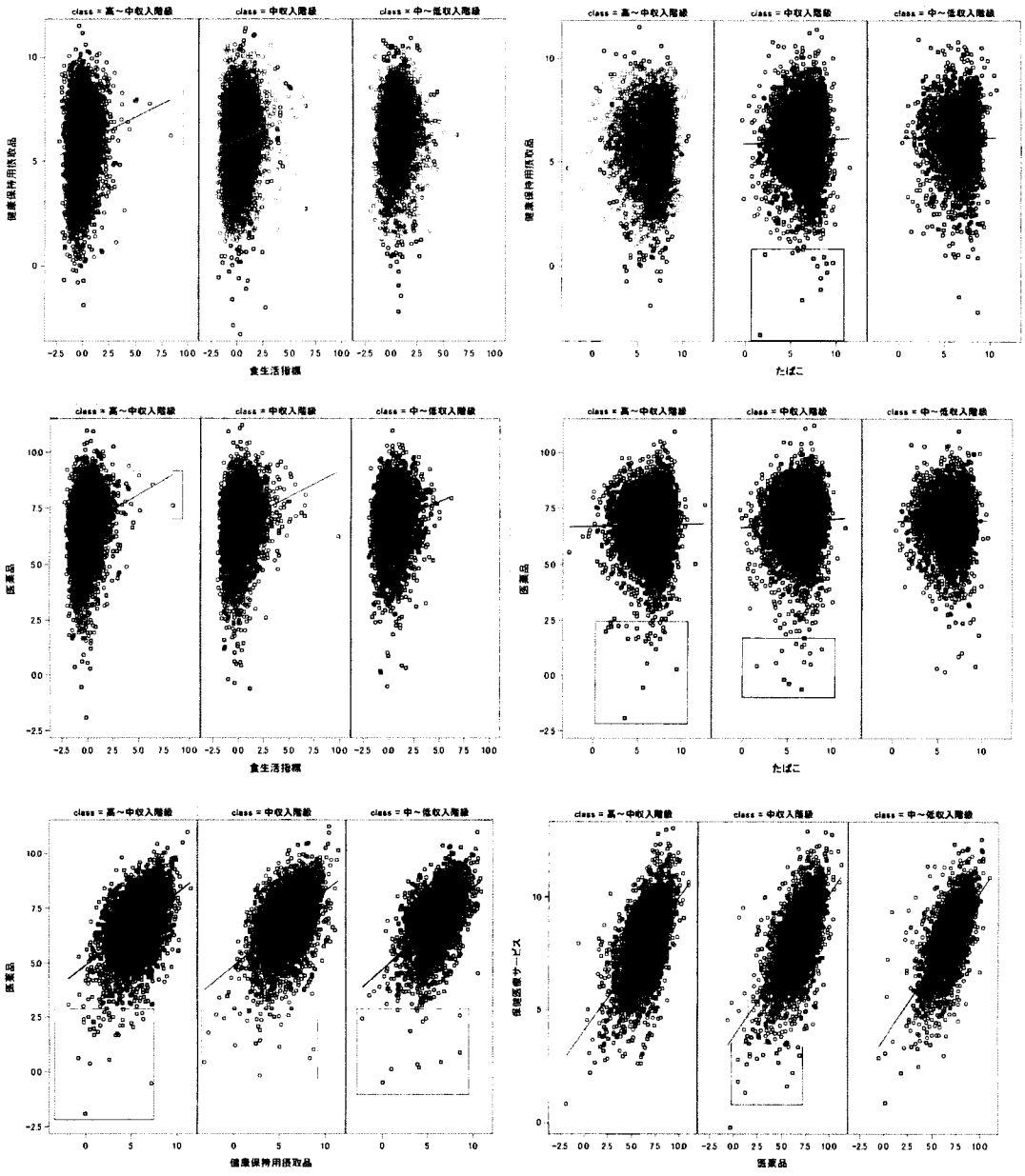
食生活と医薬品の関係では、食生活指標の得点に比べ、低収入と最高収入階級は医薬品への支出が低く、高収入階級は高い。機能性補助食品と医薬品の関係では、機能性補助食品の支出傾向に比べ、低収入層と最高収入階級は医薬品への支出が低く、高収入階級は特に高い。医薬品と保健医療サービスの支出の関係では、医薬品の支出傾向に比べ、低収入と最高収入階級は保健医療サービスへの支出が高く、その他の群は低い。特に、最高収入階級と高収入階級は医薬品への支出額は近いが、保健医療サービスの額に開きがある。尚、最低収入階級については、生活保護世帯が含まれる可能性があり、医療費が過小評価されている可能性がある。また、医薬品と保健医療サービスでは、負の相関が望ましいと考えていたが、正の相関であった。しかしこれは健康に関するイベントの違いで、医薬品も活用するが、保健医療サービスも利用するという可能性もあり、この結果だけでは OTC 医薬品が医療費削減に寄与していないとは言えない。

以上より、低収入階級において食生活指標に比して健康保持用摂取品への支出が高いが、OTC 医薬品への支出は低く、OTC 医薬品の支出傾向に比して保健医療サービスへの支出が高いという、セルフメディケーションが良好でない可能性が見られた。一方、高収入階級において食生活指標に比して OTC 医

[5]個票から算出した相関係数でも、有意性はないながら、たばこの項目は負の相関、その他の項目は正の相関があった。

薬品への支出が高く、OTC 医薬品の支出傾向に比して保健医療サービスへの支出は低いという、セルフメディケーションが良好な可能性が見られた。このように、低収入層と高収入層というセグメントでは、セルフメディケーションに対する行動が違うことが浮き彫りとなった。次に、中収入層については個票にドリルダウンし、特徴を見出す。

4-3　個票の散布図



4-4　中所得層の考察

中所得層については、個票の散布図により検討した。食生活指標が集団に比べ相当程度高い集団でも、

機能性補助食品と OTC 医薬品の使用は平均的であった。この集団に効果的な使用方法を教育することによって、より健康が高まることが期待される。たばこについては、機能性補助食品や OTC 医薬品への支出が低い集団の中でたばこへの支出額はばらつきがある。健康保持用摂取品と医薬品についても同様に、医薬品への支出が低い集団の中で健康保持用摂取品への支出額はばらつきがある。医薬品と保健医療サービスとの関係では、医薬品の使用が低い群の中に、医療費が平均（回帰直線）より低い群と高い群に分けられた。後者は、セルフメディケーション啓発の対象となり得ると言える。

5. 結語

セルフメディケーションの概念に基づき、各世帯における生活習慣、OTC 医薬品の使用、医療費の関係について考察した。

まず、収入階級別にみると、健康的な食生活の上に機能性補助食品を活用し、さらに OTC 医薬品が活用されていることが示唆された。また、収入が高いほどこれらへの支出が高い傾向も見られた。そして、低収入階級においてセルフメディケーションが良好でなく、高収入階級においてセルフメディケーションが良好な可能性が見られた。このように、セルフメディケーションの成否に、所得水準が関係することが示唆された。所得格差と健康格差の一因に、ヘルスリテラシーが挙げられている [4]。低所得層にはセルフメディケーションを含め、健康教育が必要であろう。一方、最高収入階級についても、一部セルフメディケーションが良好でない可能性が見られた。この階級は、年間収入 2,000 万円以上の 157 名という大変特異な集団である。そのため、例えば保健医療サービスに高額な自費診療が含まれているなど、特殊ケースが混在している可能性はある。しかし、国民医療費の削減の観点から考えると、この階級にも軽度な身体の不調は自分で手当てするという、セルフメディケーション啓発は必要である。

中所得層については、食生活が良好な集団に、健康保持用摂取品や医薬品の効果的な使用方法を教育する必要性が示唆された。また、医薬品の使用が低い群の中に、医療費が高めの群が見られた。この集団は、セルフメディケーション啓発の対象となり得る。

研究の限界について 2 点挙げる。まず、医薬品と保健医療サービスの項目について、筆者が想定している項目外の要因が影響していた点である。医薬品には院外処方が含まれ、保健医療サービスは世帯によって負担割合が異なる。次に、生活習慣として運動に関する項目が取り上げられなった点である。運動は健康の重要なファクターでありながら、個人差が大きい。これを変数に加えることによって、より特徴を見出すことが期待される。

本研究により、世帯の家計状況からセルフメディケーションの実態を把握し得る可能性が示唆された。今後は、他の属性でも確認することで、セルフメディケーションの啓発が必要な世帯の特徴を浮き彫りにすることを課題とする。それにより、セルフメディケーション、OTC 医薬品を活用した医療費削減施策への一資料とする。

本研究には、「擬似ミクロデータ（平成16年全国消費実態調査）」（独立行政法人 統計センター）を利用した。

引用・参考文献

1) 日本 OTC 医薬品協会　http://www.jsmi.jp/index.html（アクセス日：2014 年 5 月 14 日）
2) 総務省統計局　平成 16 年全国消費実態調査　用語の解説
http://www.stat.go.jp/data/zensho/2004/kaisetsu.htm（アクセス日：2014 年 5 月 14 日）
3) 厚生労働省　平成 22 年国民健康・栄養調査結果の概要
http://www.mhlw.go.jp/stf/houdou/2r98520000020qbb.html（アクセス日:2014 年 5 月 21 日）
4) 日本学術会議　わが国の健康の社会格差の現状理解とその改善に向けて
http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-21-t133-7.pdf（アクセス日：2014 年 5 月 26 日）

第二回 ミクロデータ分析コンテスト

規定課題

2014年5月28日

クルーズ株式会社
CROOZ
吉田大祐　長谷健司
金磊　野口光伸

参加カテゴリー：
「C）SAS または JMP の使用歴 2 年未満」

下記に SAS Enterprise Guide 6.1 を用いた「擬似ミクロデータ（平成 16 年全国消費実態調査）」の再現結果を図示する。（データの可視化については Microsoft Excel 2010 を使用している。）

1.「第1−1表 集計世帯数」の再現結果

（各レコードを、単純にカウントしたもの）

		世帯人口									
		2人	3人	4人	5人	6人	7人	8人	9人	10人	合計
有業人員	1人	4,124	2,806	4,132	1,436	256	51	6	0	0	12,913
	2人	3,239	5,391	4,281	1,943	494	162	29	0	0	15,459
	3人	0	1,035	1,031	559	232	84	6	3	0	2,950
	4人	0	0	324	220	104	31	12	0	0	691
	5人	0	0	0	27	6	7	0	0	0	40
	6人	0	0	0	0	3	3	0	0	0	6
	不詳	75	203	256	220	119	52	28	12	3	968
	合計	7,438	9,537	9,944	4,405	1,214	390	81	15	3	32,027

2.「第1−2表 世帯数分布」の再現結果

（各レコードを、集計用乗率で重み付けして、カウントしたもの）

		世帯人口									
		2人	3人	4人	5人	6人	7人	8人	9人	10人	合計
有業人員	1人	64,691	61,284	86,290	22,740	3,813	831	84	0	0	219,743
	2人	50,138	91,523	64,868	29,616	6,801	2,336	441	0	0	205,723
	3人	0	15,912	15,615	7,963	3,302	1,137	76	36	0	44,041
	4人	0	0	4,851	3,345	1,354	422	195	0	0	10,166
	5人	0	0	0	383	80	108	0	0	0	570
	6人	0	0	0	0	49	51	0	0	0	99
	不詳	1,006	3,287	4,216	3,519	1,761	727	401	170	33	15,120
	合計	115,835	132,005	155,850	67,565	17,161	5,611	1,197	207	33	495,465

3.「第1−3表　世帯数分布（10万分比）」の再現結果

		世帯人口									
		2人	3人	4人	5人	6人	7人	8人	9人	10人	合計
有業人員	1人	13,057	12,369	13,381	4,590	770	168	17	0	0	44,351
	2人	10,119	10,399	13,092	5,977	1,373	471	89	0	0	41,521
	3人	0	3,211	3,152	1,607	666	229	15	7	0	8,889
	4人	0	0	979	675	273	85	39	0	0	2,052
	5人	0	0	0	77	16	22	0	0	0	115
	6人	0	0	0	0	10	10	0	0	0	20
	不詳	203	664	851	710	355	147	81	34	7	3,052
	合計	23,379	26,643	31,455	13,637	3,464	1,132	242	42	7	100,000

4.「第2表　支出（消費支出 及び 十大費目）」の再現結果

		集計世帯数	世帯数分布（抽出率調整済）
世帯人員が4人の有業人員	1人	4,132	66,299
	2人	4,201	64,868
	3人	1,031	15,615
	4人	324	4,851
	不詳	256	4,216
	合計	9,944	155,850
	1人又は2人	8,333	131,168

		消費支出										
		合計	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
世帯人員が4人の有業人員	1人	¥305,234	¥71,543	¥17,556	¥18,854	¥8,383	¥13,579	¥11,656	¥42,703	¥21,202	¥31,959	¥57,801
	2人	¥347,740	¥78,472	¥12,932	¥20,621	¥8,917	¥14,876	¥10,528	¥52,141	¥29,485	¥23,681	¥76,078
	3人	¥380,521	¥83,796	¥12,583	¥23,045	¥10,276	¥16,561	¥10,331	¥52,406	¥22,513	¥27,852	¥121,160
	4人	¥399,962	¥85,083	¥19,500	¥22,490	¥11,763	¥14,096	¥10,812	¥51,310	¥4,809	¥32,769	¥148,428
	不詳	¥379,882	¥82,134	¥24,296	¥22,238	¥7,843	¥14,238	¥10,011	¥43,521	¥49,453	¥31,206	¥94,942
	合計平均	¥335,438	¥76,362	¥15,345	¥20,214	¥8,685	¥14,452	¥10,982	¥47,894	¥32,442	¥32,269	¥75,588
	1人又は2人	¥326,256	¥74,970	¥15,269	¥19,728	¥8,647	¥14,221	¥11,103	¥47,371	¥35,298	¥32,810	¥66,839

以上、「擬似ミクロデータ（平成 16 年全国消費実態調査）」について、独立行政法人統計センターの公表値と再現一致を確認した。

2014年5月28日
クルーズ株式会社
吉田大祐　長谷健司
金磊　野口光伸

参加カテゴリー：
「C）SASまたはJMPの使用歴2年未満」

自由課題テーマ

「ゲーム・インターネット好きの世帯は、どのような食生活をしているのか？？」

はじめに

弊社クルーズ株式会社（http://crooz.co.jp/）は、SAS Institute Japan様と同じく六本木ヒルズ森タワーに所在しており、ソーシャルゲームやネット通販を中心に世界中にインターネットサービスを提供するエンターテインメント企業です。現在データマイニング分野の業務強化を行っており、今回の「ミクロデータ分析コンテスト」を通して、社内データアナリストのスキル向上とSASプロダクトを弊社に導入した際の将来の可能性を導き出したいと考えて参加させて頂いています。ちなみに弊社データアナリストは全員SASプロダクトの業務経験はなく、このコンテストで初めて使用します。

またご了承いただきたい点として、弊社の企業コンセプトは「**“オモシロカッコイイ”をツクル**」であり、他の参加チームの方々とは書類の形式が若干異なるかと思いますが、楽しんで読んでいただければ幸いです。データサイエンティストの不足が課題となっている中、難しいと思われがちな統計や分析の世界を、より多くの人々が楽しく魅力的なものに感じ、データサイエンティストの仲間となり、今後も成長し続けるビックデータによる情報社会を進化・革新を期待しています。

1. テーマの選定理由

弊社で特に注力しているソーシャルゲーム分野事業において、通常はゲームユーザーの方々と直接お会いすることはないため、どういった生活・消費行動の傾向があるのかを「擬似ミクロデータ」より調べたい思いこのテーマを選定した。生活・消費行動の傾向から新たなマーケティング施策へも展開できることを行いたい。

2. 前提条件

- データ加工・分析にはSAS Enterprise Guide 6.1およびSAS Enterprise Miner Workstation 13.1を用いる。
- 独立行政法人統計センター「擬似ミクロデータ（平成16年全国消費実態調査）」の次の図表1～3のデータ列を利用する。今回は食品支出の傾向を分析するが、それが世帯や収入で影響を及ぼしていないかを念のため確認をする。

図表 1．目的変数

「ゲーム・インターネット好き」を定義する項目
教養娯楽用品
通信

図表 2．説明変数 1

世帯、年収に関する項目
世帯人員
有業人員
年間収入 (1,000円単位のため1,000倍する)

図表 3．説明変数 2

食品支出に関する項目				
食料	魚肉練製品	野菜・海藻	油脂	他の飲料
穀類	他の魚介加工品	生鮮野菜	調味料	酒類
米	肉類	乾物・海藻	菓子類	外食
パン	生鮮肉	大豆加工品	調理食品	一般外食
めん類	加工肉	他の野菜・海藻加工品	主食的調理食品	学校給食
他の穀類	乳卵類	果物	他の調理食品	
魚介類	牛乳	生鮮果物	飲料	
生鮮魚介	乳製品	果物加工品	茶類	
塩干魚介	卵	油脂・調味料	コーヒー・ココア	

「ゲーム・インターネット好き」の定義は、「教養娯楽用品」および「通信」の支出項目の合計金額が 18,000 円以上の世帯とする。

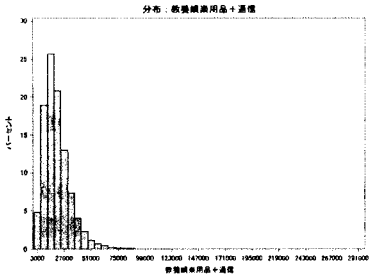
- 「教養娯楽用品」には「テレビゲーム機，ゲームソフト等，他のがん具」の他に「文房具」および「運動用具類」も含まれる（総務省統計局より）が、支出金額が多いほど「ゲーム好き」と仮定する。
- 「通信」については平成 16 年のデータであるため、現在の通信環境とは大きく異なるが、支出金額が多いほど「インターネット好き」と仮定する。
- 18,000 円という値は、「教養娯楽用品」と「通信」の合計金額の中央値（18,111 円）から決定した（図表４）。

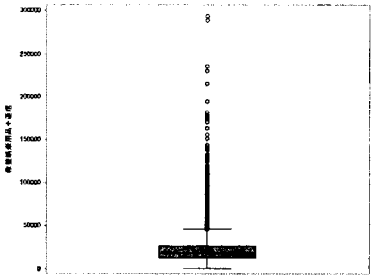
図表４．「教養娯楽用品」と「通信」の合計値の要約統計量

分析変数：教養娯楽用品＋通信							
平均	標準偏差	最小値	最大値	N	下側四分位点	中央値	上側四分位点
20579.09	12826.78	0	293154.42	32027	12299.61	18111.24	25818.65

弊社のような Web 業界でのデータアナリストは、高度あるいは精度の高い分析モデルよりも、刻々と変化する環境の中で最短で効果的な次のアクションを起こせるマーケティング分析が必要とされることが多く、最終意思決定の担当者でも理解しやすいディシジョンツリーによる分析を行う。

図表５．【参考】「教養娯楽用品」と「通信」の合計値の「ピクトグラム」および「箱ひげ図」





3. データマイニング

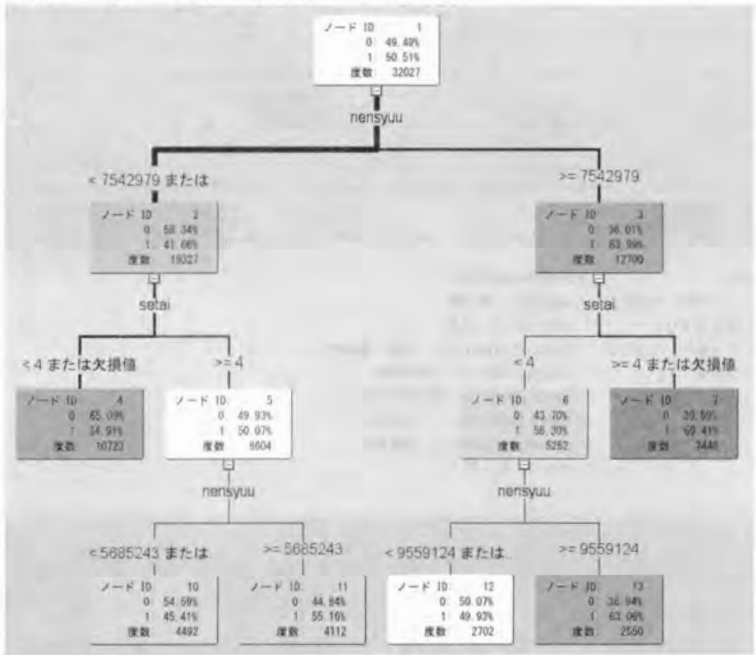
「教養娯楽用品」および「通信」の支出項目の合計金額が 18,000 円以上の世帯を「1」、18,000 円未満を「0」としてターゲットのフラグを作成し、ディシジョンツリーで分析する。

① 世帯・有業人員、年間収入と関係はあるか？

まず、世帯・有業人員や年間収入の関連性はあるのかを「説明変数 1」を用いてディシジョンツリーで分析する。

結果は図表 6 となり、最も影響のある項目は「年間収入」、次に「世帯人数」となる。「年間収入」の 754 万円で分岐した「ターゲット」が「1」の度数はほぼ同値となり、また「世帯人数」に比例し「教養娯楽用品」および「通信」も増加する傾向があるため、「ゲーム・インターネット好き」世帯は、世帯・有業人員や年間収入とは関係は低いと言える。

図表 6．「説明変数 1」によるディシジョンツリー



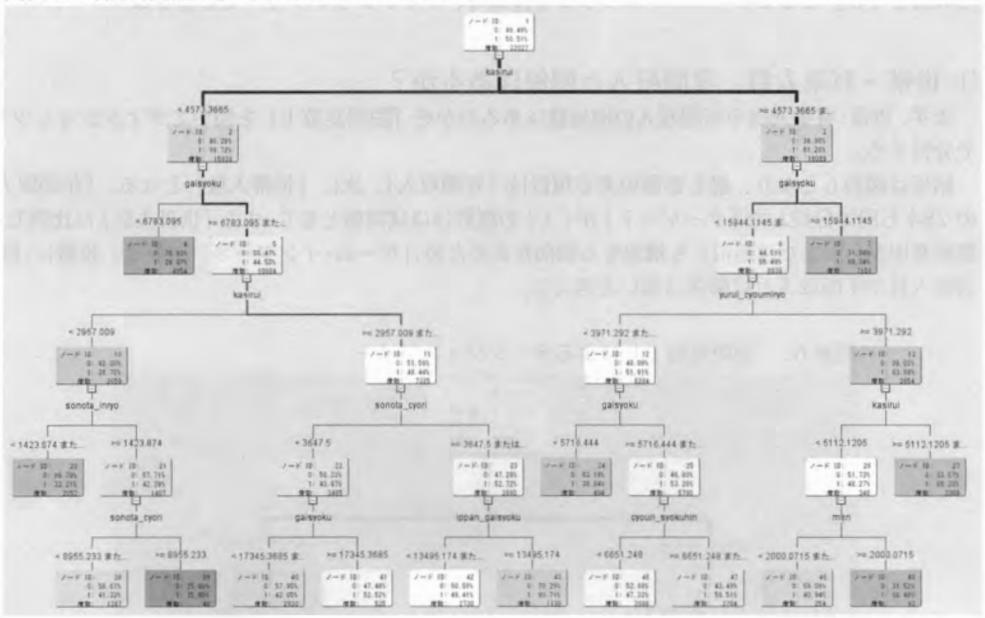
[図表内の目的変数]
・「教養娯楽用品」と「通信」の合計が 18,000 円未満の世帯を「0」
・「教養娯楽用品」と「通信」の合計が 18,000 円以上の世帯を「1」

[図表内の説明変数]
・nensyuu：　年間収入
・setai：　世帯人員

② 食生活と関係はあるか？

「説明変数２」を用いたディシジョンツリーを図表７に示す。

図表７.「説明変数１」によるディシジョンツリー



[図表内の目的変数]
- 「教養娯楽用品」と「通信」の合計が18,000 円未満の世帯を「０」
- 「教養娯楽用品」と「通信」の合計が18,000 円以上の世帯を「１」

[図表内の説明変数]
- kasirui：　菓子類
- gaisyoku：　外食
- yurui_cyoumiryo：　油脂・調味料
- sonota_inryou：　他の飲料
- sonot_cyouri：　他の調理食品
- ippan_gaisyoku：　一般外食
- cyouri_syokuhin：　調理食品
- men：　めん類

傾向としては
「ゲーム・インターネット好き」は…

⇩

「お菓子」をよく食べる！
「外食」によく行く！

…やっぱりそうか！

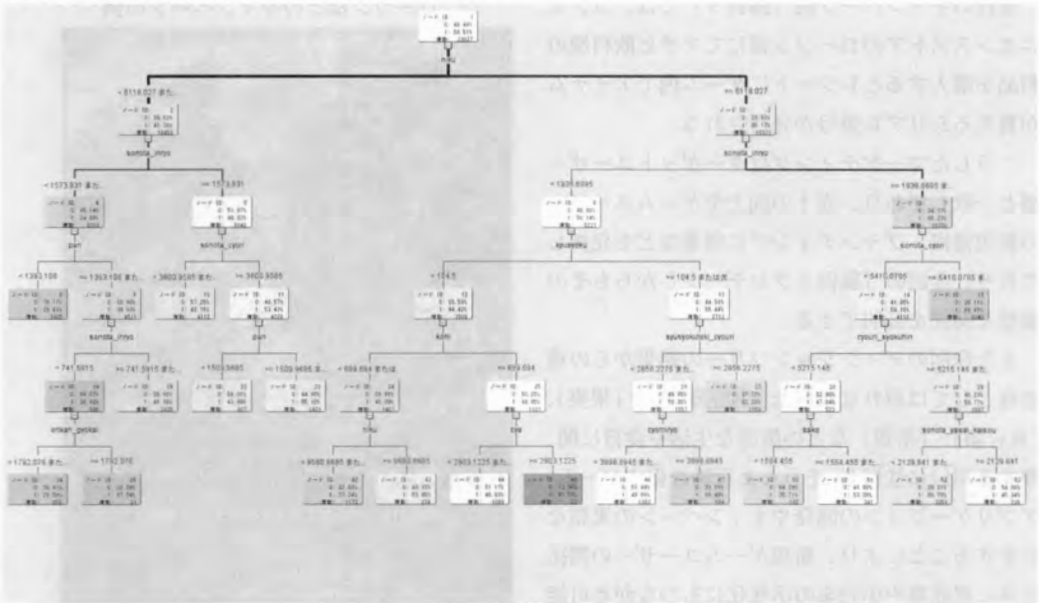
料理よりもゲーム・インターネットをする時間の方が大切！

4

③　② から「菓子類」と「外食」の項目を除き、他に大きな要因はあるか？

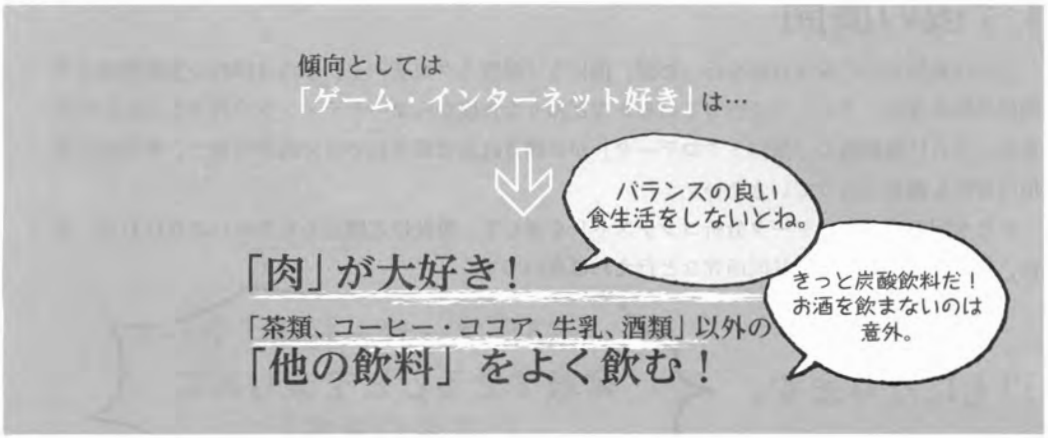
「菓子類」と「外食」以外の他の説明変数では、どういった傾向になるのか、「説明変数２」から「菓子類」と「外食」を除いてディシジョンツリーを分析する。結果を図表８に示す。

図表８．「説明変数１」によるディシジョンツリー



[図表内の目的変数]
・「教養娯楽用品」と「通信」の合計が18,000円未満の世帯を「0」
・「教養娯楽用品」と「通信」の合計が18,000円以上の世帯を「1」

[図表内の説明変数]
・niku：　肉類
・sonota_inryou：　他の飲料
・pan：　パン
・sonot_cyouri：　他の調理食品
・kyusyoku：　学校給食
・kohi：　コーヒー・ココア
・syusyokuteki_cyouri：主食的調理食品
・cyouri_syokuhin：　調理食品
・enkan_gyokai：　塩干魚介
・cya：　茶類
・cyomiryo：　調味料
・sake：　酒類
・sonota_yasai_kaisou：他の野菜・海藻加工品



4.アクションプラン

既に弊社を含めソーシャルゲーム業界では、菓子メーカーや清涼飲料メーカーとのコラボレーション商品の開発やキャンペーンを実施しており、また外食チェーンレストランとのタイアップも多く存在している。

弊社のキャンペーン例（図表9）では、コンビニエンスストアのローソン様にてアサヒ飲料様の製品を購入するとレシートにゲーム内でアイテムが貰えるシリアル番号が発行される。

こうしたマーケティングはターゲットユーザー層と一致しており、売上の向上やゲームユーザーの新規獲得、ブランディングに効果などを発揮しており、今回の「擬似ミクロデータ」からもその重要な関係を証明できる。

また今回のディシジョンツリーの結果からの重要度としては現れなかった「生鮮野菜」や「果実」、「魚介類」、「茶類」などの健康な生活や食育に関連する項目に注目し、そうした行動を促すゲーム・アプリケーションの開発やキャンペーンの実施などをすることにより、新規ゲームユーザーの開拓と共に農産業や小売業の活性化にもつながる可能性はある。

図表9. 弊社のソーシャルゲームと、アサヒ飲料様、ローソン様とのキャンペーンの例

今後の展開

今回は食品支出に絞り分析を行ったが、他にも「擬似ミクロデータ」からは詳細な世帯環境との関係や医療支出、スポーツを行っているかなど様々な角度からマーケティング分析ができる活用できる。さらに最新版の「擬似ミクロデータ」が公開されれば時系列での分析が可能で、世帯での環境の変化も調査を行っていきたい。

また今回の「ミクロデータ分析コンテスト」を通して、弊社にご関心をお寄せいただければ、今後、是非とも情報交換や共同研究など行えれば幸いです。

以上になります。

「SAS ユーザー総会」で
発表できることを楽しみに
しております！

6

Let's データ分析 第2回ミクロデータ分析コンテスト：規定課題

KG あならいず
笹谷知輝　大野嵩護

カテゴリー C

要旨

本稿では、本コンテストの規定課題である「集計結果」の再現の過程および結果を示す。なお、使用するデータは「平成16年全国消費実態調査」の個票データに基づいて、独立行政法人統計センターによって作成された「教育用疑似ミクロデータ」である。これらのデータを取り込む際、DOSで7つのcsvファイルをひとつのファイルに結合した後、importプロシジャを用いてSASにインポートした。

キーワード：教育用疑似ミクロデータ　集計用乗率

1．第1-1表、第1-2表、第1-3表の作成

第1-1表、第1-2表、第1-3表はfreqプロシジャを用いて作成した。第1-2はweightステートメントを用いて集計用乗率を加味した。また、第1-3では、10万分比にするために、10^5 / 全データ数、すなわち、10^5 / 495465と集計用乗率の積からなる10万分比集計用乗率を作成し重みづけを行った。

第1－1表　集計世帯数

```
proc freq data=kadai1;
    table ShuugyouJinin*SetaiJinin
        / missing nopercent nocol norow;
    format ShuugyouJinin ShuugyouJinin.
        SetaiJinin SetaiJinin.;
run;
```

表：ShuugyouJinin * SetaiJinin

ShuugyouJinin(有業人員)	SetaiJinin(世帯人員) 2人	3人	4人	5人	6人	7人	8人	9人	10人	合計
不詳	75	203	256	220	119	52	28	12	3	968
1人	4124	3908	4132	1436	256	51	6	0	0	13913
2人	3239	3391	4201	1943	494	162	29	0	0	13459
3人	0	1035	1031	559	232	84	6	3	0	2950
4人	0	0	324	220	104	31	12	0	0	691
5人	0	0	0	27	6	7	0	0	0	40
6人	0	0	0	0	3	3	0	0	0	6
合計	7438	8537	9944	4405	1214	390	81	15	3	32027

第1－2表　世帯数分布

```
proc freq data=kadai1;
    table workmember*HHmember
        / out=kadai1_2.kadai1_2 missing
        nopercent nocol norow format=F7.0;
    weight weight;
    format workmember workmember.
        HHmember HHmember.;
run;
```

表：workmember * HHmember

workmember(有業人員)	HHmember(世帯人員) 2人	3人	4人	5人	6人	7人	8人	9人	10人	合計
不詳	1006	3287	4216	3519	1761	727	401	170	33	15120
1人	64691	61284	66299	22740	3813	831	84	0	0	219743
2人	50138	51523	64868	29616	6801	2336	441	0	0	205723
3人	0	15912	15515	7963	3302	1137	76	36	0	44041
4人	0	0	4851	3345	1354	422	195	0	0	10168
5人	0	0	0	383	80	108	0	0	0	570
6人	0	0	0	0	49	51	0	0	0	99
合計	115835	132005	155850	67565	17161	5611	1197	207	33	495465

第1－3表　世帯数分布（10万分比）

```
proc freq data=kadai1;
    table workmember*HHmember
        / out=kadai1_3 missing nopercent
        nocol norow format=f7.0;
    weight weight2;
    format workmember workmember.
        HHmember HHmember.;
run;
```

表：workmember * HHmember

workmember(有業人員)	HHmember(世帯人員) 2人	3人	4人	5人	6人	7人	8人	9人	10人	合計
不詳	203	664	851	710	355	147	81	34	7	3052
1人	13057	12369	13381	4590	770	168	17	0	0	44351
2人	10119	10399	13092	5977	1373	471	89	0	0	41521
3人	0	3211	3152	1607	666	229	15	7	0	8889
4人	0	0	979	675	273	85	39	0	0	2052
5人	0	0	0	77	16	22	0	0	0	115
6人	0	0	0	0	10	10	0	0	0	20
合計	23379	26643	31455	13637	3464	1132	242	42	7	100000

２．第２表の作成

第２表については、総数の１行、世帯人員が４人（特掲行を除く）の６行をそれぞれ tabulate プロシジャで作成し、特掲の１行は世帯人員が４人で有業人員が１人と２人の和を先述の６行の表から計算し、set ステートメントを使ってそれらを結合することで作成した。

第２表

```
/*世帯人員が4人で有業人員が1-4およびその総数*/
proc tabulate
    data=kadai2_original out=hh4_w1to4;
    where HHmember=4;
    class HHmember workmember / missing;
    var weightx consume food house
        elec_gas furniture
        clothes medical transport edu amuse
        other;
    table HHmember*(workmember all),
        N consume*mean*f=8.0
        food*mean*f=8.0 house*mean*f=8.0
         elec_gas*mean*f=8.0 furniture*mean*f=8.0
        clothes*mean*f=8.0 medical*mean*f=8.0
        transport*mean*f=8.0 edu*mean*f=8.0
        amuse*mean*f=8.0 other*mean*f=8.0;
    weight weightx;
    keylabel all="総数" mean=" " N="集計世帯数";
run;

/*総数（世帯人員の制限なし）*/
proc tabulate data=kadai2_original out=hhtotal;
    class / missing;
    var consume food house elec_gas furniture
        clothes medical
        transport edu amuse other;
    table N consume*mean*f=8.0
        food*mean*f=8.0 house*mean*f=8.0
        elec_gas*mean*f=8.0
        furniture*mean*f=8.0
        clothes*mean*f=8.0 medical*mean*f=8.0
        transport*mean*f=8.0 edu*mean*f=8.0
        amuse*mean*f=8.0 other*mean*f=8.0;
    weight weightx;
    keylabel all="総数" mean=" " N="集計世帯数";
run;
```

```
/*（特掲）１人または２人*/
data tokkei_ave;
    array expense {11}
        consume_Mean food_Mean
        house_Mean elec_gas_Mean
        furniture_Mean
        clothes_Mean medical_Mean
        transport_Mean edu_Mean
        amuse_Mean other_Mean;
    array total {11};
    array result {11};
    set end=final_obs;

    do i=1 to 11;
        total{i}+sum_weight*expense{i};
    end;
    one_two+sum_weight;
    if final_obs then
        do;
         do i=1 to 11;
           result{i}=total{i} / one_two;
         end;
         output;
        end;
run;
```

OBS		集計世帯数	世帯数分布	消費支出	食料	住居	光熱・水道	家具・家事用品	被服および履物	保健医療	交通・通信	教育	教養娯楽	その他の消費支出
1	総数	32027	495465	328140	72883	17687	19238	9204	14138	11366	47961	22270	31389	82003
2	うち世帯人員が4人	9944	4216	335438	76362	15345	20214	8885	14452	10987	47894	33442	32269	75588
3	1人	4132	66299	305234	71543	17556	18854	8383	13579	11656	42703	31202	31959	57801
4	2人	4201	64868	347740	78472	12932	20621	8917	14876	10538	52141	39485	33681	76078
5	3人	1031	15615	380521	83796	12583	23045	10276	16561	10331	52406	22513	27852	121160
6	4人	324	4851	399962	85083	18500	22490	11763	14096	10812	51310	4509	32769	148628
7	不詳	256	155850	379862	82134	24296	22238	7843	14238	10011	43521	49453	31206	94942
8	（特掲）1人又は2人	8333	131168	326256	74970	15269	19728	8647	14221	11103	47371	35298	32810	66839

年代別の魚食傾向に関する考察

—教育用疑似ミクロデータを用いて—

KG あならいず[1]
笹谷 知輝[2]　　大野 嵩護[3]
カテゴリー C

Analyzing Fish-Eating Tendency by Age Group

— Using the Pseudo Micro Data for Educational Purpose —

Tomoki Sasaya[2]　　Shugo Ono[3]
Kwansei Gakuin University

要旨

昨今、日本では消費傾向の変化による「魚離れ」が問題となっている。そこで、疑似ミクロデータを用いて、年代別の魚介類支出の割合を算出し、他の品目と比較することで「若者の魚離れ」の実態を検証する。

キーワード：教育用疑似ミクロデータ、魚離れ、若者、ファストフィッシュ（Fast Fish）

1．はじめに

国連食糧農業機関によると、日本の魚介類産出量は世界で 5 番目の規模[1]を誇り、日本の一人あたりの魚介類消費量は世界で 6 番目[2]であることから、日本は魚と密接な関係を持っているといえる。しかし、近年「若者の魚離れ」が指摘されている[3]。「魚離れ」の原因として、魚の下処理や調理器具が汚れることなど、多忙な現代人のライフスタイルにそぐわないことが挙げられる[3]。これは日本の水産業界にとって大きな問題であるため、様々な解決策が講じられている。そのひとつが、ファストフィッシュ（Fast Fish）商品の選定である。ファストフィッシュとは「手軽・気軽においしく、水産物を食べること及びそれを可能にする商品や食べ方のことで、今後普及の可能性を有し、水産物の消費拡大に資するもの」と水産庁は定義している[4]。このように、「若者の魚離れ」は社会から一定の関心を集めている。そこで若者の魚離れの実態に関して、平成 16 年全国消費実態調査のデータから作成された教育用擬似ミクロデータ（以後「ミクロデータ」と呼称する）を用いて分析しその結果について考察する。

本稿では、年齢階級ごとに各食料品目の平均支出を比較し、「若者の魚離れ」を概観する。食料支出に占める魚介類支出の割合の年齢階級による違いを明らかにし、魚介類と他の食料品目を比較した上で、魚介類の代わりに消費していると考えられる食料品目を検討する。佐藤ら[5]は、魚の嗜好は親等の調理担当者や家庭での食生活が大いに関係していると結論付けている。したがって、世帯主の年齢階級別に魚食の動向を考察することは正当であるといえる。

---

[1] 関西学院大学共通教育センター 「データ分析研究会（KG あならいず）」
[2] 関西学院大学経済学部 3 年
[3] 関西学院大学商学部 3 年

2．年齢階級別でみる食料支出と魚介類支出の検討

魚介類の支出を検討するにあたり、ミクロデータから食料品目の支出を考察した。「食料」に大分類されている食料品目のうち、「穀類」、「魚介類」、「肉類」、「乳卵類」、「野菜・海藻」を「主食・副食」グループに、「菓子類」、「果物」、「油脂・調味料」、「飲料」、「酒類」を「嗜好品」グループに、「調理食品」、「外食」を「その他」グループとして、三つのグループ分けを行った。なお、世帯の食料費の支出は世帯人員が増えれば当然増加するので、まず、食料支出と世帯人員の相関係数を求めた。P 値<0.0001 で相関係数が 0.273 の弱い正の相関があったので、世帯主の年齢階級幅を 5 歳とした年齢階級別に一人当り平均支出額を、乗率による重み付けを行った上で算出した。また、世帯主の年齢階級が不詳の世帯は除外した。

乗率を換算した世帯主の年齢階級のサンプル分布を図 1 に示す。サンプル数の少ない「24 歳未満」、「25～29 歳」の年齢階級を「29 歳未満」に、また「65～69 歳」、「70～74 歳」、「75 歳以上」の年齢階級を「65 歳以上」に再分類した（図 2)。再分類後の年齢階級は「29 歳未満」を 1、「30～34 歳」を 2、以降を 9 までの整数値とした後、分析を行った。

まず、各食品グループの各世帯の一人当たりの平均支出額 y と、世帯主の年齢階級 x の関係を調べるために次の回帰式 (1) を用いた。

$$y = \alpha + \beta x \qquad (1)$$

計算結果を表 1 に示す。表 1 より、「主食・副食」グループのβが、他のグループの結果より大きいことがわかる。したがって、「主食・副食」グループが世帯主の年齢に最も影響を受けているといえる。次に、世帯主の年齢が低いほど「魚介類」の支出額が少なくなるかどうかを調べる。上記 (1) 式を用いて、「魚介類」の支出額を「主食・副食」グループの支出額の合計で割った値を y に、世帯主の年齢階級を x として回帰分析を行った。また、「主食・副食」グループ内の「魚介類」以外の食品項目である「穀類」、「肉類」、「乳卵類」、「野菜・海藻」についても同様の回帰分析を行い、その結果を比較した（表 2)。これらの食料品目の中では「魚介類」の変化量が最も大きく、魚介類の支出額の変化と世帯主の年齢階級の変化に正の相関関係があることがわかる。すなわち、世帯主の年齢階級が低い世帯ほど「魚介類」への支出が少ないといえる。

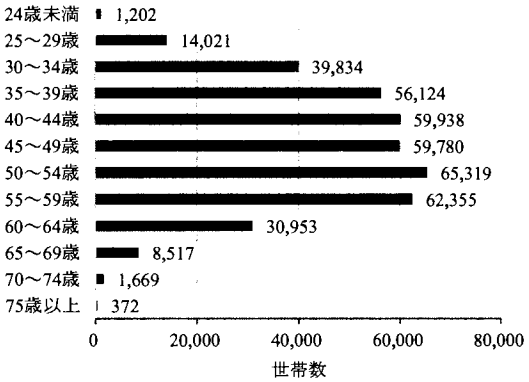


図 1　世帯主年齢階級別の分布

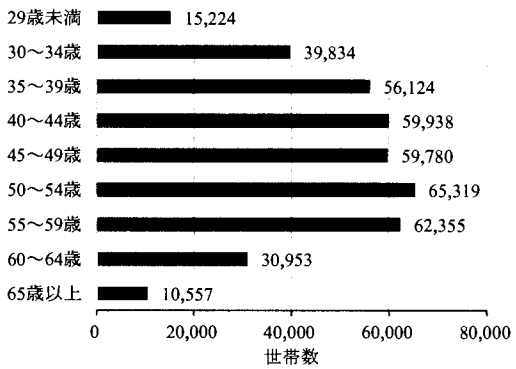


図 2　世帯主年齢階級別の分布（再分類後）

表 1　各食品グループと年齢階級の回帰分析

食品グループ	α	β	決定係数
主食・副食	3,617.7	1,320.9	0.33
嗜好品	3,030.9	497.3	0.15
その他	6,178.6	92.9	0.00

表 2「主食・副食」グループの各項目と年齢階級の回帰分析結果

食品項目	α	β	決定係数
魚介類	0.119	0.0157	0.34
野菜・海藻	0.235	0.0066	0.08
穀類	0.247	-0.0036	0.02
乳卵類	0.151	-0.0088	0.21
肉類	0.248	-0.0098	0.17

3．全食料支出における「魚介類」の支出額

「魚介類」の支出は世帯主の年齢が低いほど少なくなることがわかった。次に、若年層が「魚介類」の代わりに消費している食料品目について検討した。まず、被説明変数を全食料支出に占める各食料品目の支出の割合とし、世帯主の年齢階級を説明変数として回帰分析を行った。その結果得られた各回帰直線の回帰係数の比較（図 3）から、年齢階級別の各食料品目の全食料支出に占める割合の違いがわかる。「魚介類」は他の食料品目と比べて変化の割合が最も大きい。一方、「外食」は他の食料品目と比べて変化量が大きい負の数になっている。これは、世帯主の年齢階級が低いほど食料支出に占める「魚介類」の割合が小さく、食料支出に占める「外食」の割合が大きくなることを示している。図 4 は年齢階級別の全食料品目の支出額の積上げグラフである。食料の支出額は年齢階級が低いほど少なく、各食料品目の支出も少なくなっている。しかし、「外食」の支出は年齢階級間ではあまり差が見られない。全食料支出に占める「外食」の割合を図 5 に、「魚介類」の割合を図 6 に示す。年齢階級が低いほど「外食」の全食料支出に占める割合は大きくなるといえる。 一方、「魚介類」の全食料の支出額に占める割合は年齢階級が低いほど小さい。そこで、全食料支出に占める「魚介類」と「外

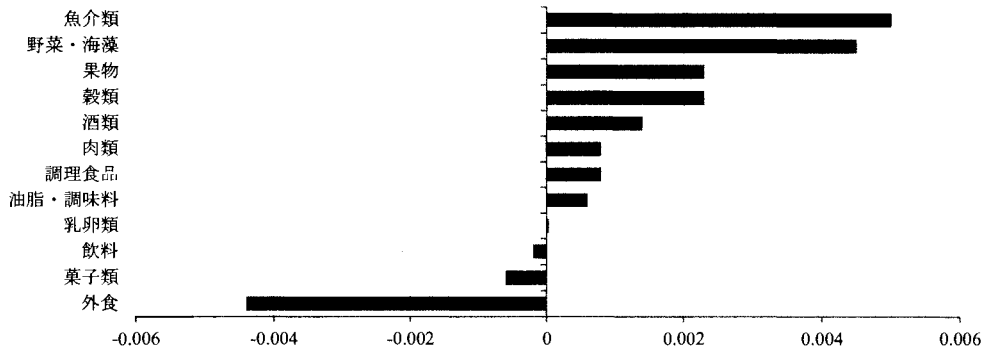


図 3　全食料品目の回帰直線の回帰係数

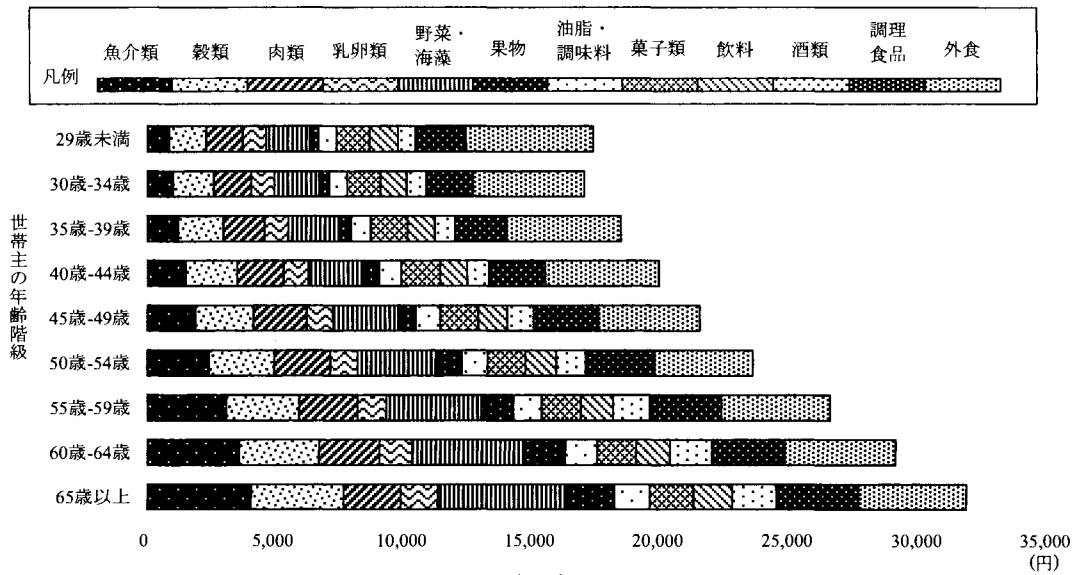


図 4 平均食料支出額

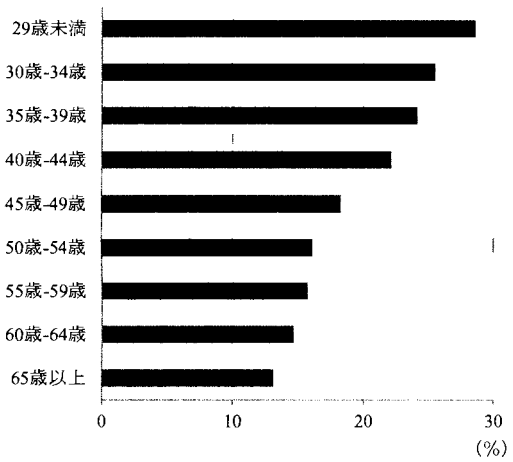


図5　一人当たりの全食料支出に占める「外食」の割合

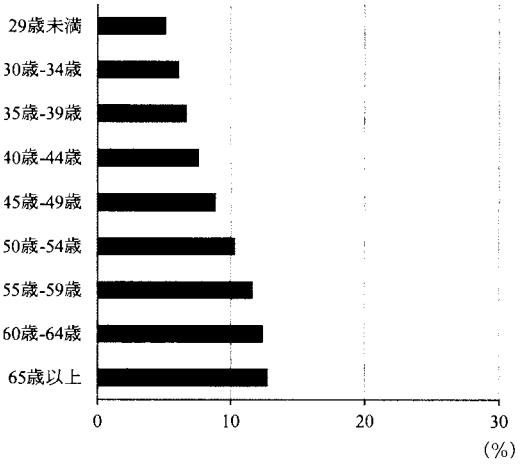


図6 一人当たりの全食料支出に占める「魚介類」の割合

食」のそれぞれの平均支出額の割合に対して、proc corr を使って無相関検定を行ったところ、相関係数が-0.5518 であり、P 値が 0.0001 未満という値が得られたため、負の相関があるといえる。

この分析の結果、全食料支出額において、世帯主の年齢が下がるにつれて「魚介類」の支出額の割合が減少することがわかった一方で、「外食」の支出額の割合は大きくなっており、世帯主の年齢階級が低いほど外食に多く支出しているといえる。これは、労働形態の変化や外食産業の発達によるライフスタイルの変化が、こうした消費傾向を後押ししていると考えられる。

4．まとめ

本稿では「平成 16 年全国消費実態調査」教育用疑似ミクロデータを用いて分析を行った。まず、食料品目を 3 グループに分類し、各食品グループと年齢階級との関係を分析したところ、「主食・副食」グループが最も年齢階級に影響を受けているという結果を得た。また、魚介類の支出額の変化と世帯主の年齢階級の変化には正の相関関係があり、世帯主の年齢階級が低いほど、「魚介類」の支出額が少ないことがわかった。そして、若者が「魚介類」の代わりに消費している食料品目を調べるために、全食料品目に占める各食料品目の支出額の割合を比較した結果、年齢が下がるほど「外食」への支出の割合が大きいことがわかった。以上より、世帯主の年齢に基づき「魚介類」に対する消費傾向を見ることができたが、世帯構成員全員の年齢を含むデータを用いることができれば、より正確な分析が可能になると考える。

参考文献

1) 農林水産省：我が国における魚介類摂取の特徴,<http://www.maff.go.jp/j/syouan/tikusui/gyokai/g_kenko/tokucyo/>（参照 2014-05-28）.
2) 水産庁：水産業をめぐる情勢の変化 <http://www.jfa.maff.go.jp/j/kikaku/kihonkeikaku/pdf/shiryo2_4.pdf>（参照 2014-05-28）.
3) 農林水産省：第２節 急速に進む「魚離れ」～魚食大国に翳り<http://www.maff.go.jp/hakusyo/sui/h18/html/s1_1_2.htm>（参照 2014-05-28）.
4) 水産庁：Fast Fish（ファストフィッシュ）関係資料 <http://www.jfa.maff.go.jp/j/kikaku/shiawase2.html>（参照 2014-05-28）.
5) 佐藤 和美，薬師寺 國人：若者の魚嗜好と魚食の実態研究，鎌倉女子大学紀要，第 10 号，pp.111-118 (2003).

規定課題

谷口 裕明
ソニー銀行㈱　総合リスク管理部
参加カテゴリー：C

基礎集計表の計数について、再現した結果を以下に示す。

第１－１表　集計世帯数（各レコードを単純にカウントしたもの）

		総数	世帯人員								
			2	3	4	5	6	7	8	9	10
総数		32,027	7,438	8,537	9,944	4,405	1,214	390	81	15	3
有業人員	1	13,913	4,124	3,908	4,132	1,436	256	51	6	0	0
	2	13,459	3,239	3,391	4,201	1,943	494	162	29	0	0
	3	2,950	0	1,035	1,031	559	232	84	6	3	0
	4	691	0	0	324	220	104	31	12	0	0
	5	40	0	0	0	27	6	7	0	0	0
	6	6	0	0	0	0	3	3	0	0	0
	VV	968	75	203	256	220	119	52	28	12	3

第１－２表　（各レコードを、集計用乗率で重み付けして、カウントしたもの）

		総数	世帯人員								
			2	3	4	5	6	7	8	9	10
総数		495,465	115,835	132,006	155,850	67,565	17,161	5,611	1,197	207	33
有業人員	1	219,743	64,691	61,284	66,299	22,740	3,813	831	84	0	0
	2	205,723	50,138	51,523	64,868	29,616	6,801	2,336	441	0	0
	3	44,041	0	15,912	15,615	7,963	3,302	1,137	76	36	0
	4	10,168	0	0	4,851	3,345	1,354	422	195	0	0
	5	570	0	0	0	383	80	108	0	0	0
	6	99	0	0	0	0	49	51	0	0	0
	VV	15,120	1,006	3,287	4,216	3,519	1,761	727	401	170	33

第１－３表　（10 万分比）

		総数	世帯人員								
			2	3	4	5	6	7	8	9	10
総数		100,000	23,379	26,643	31,455	13,637	3,464	1,132	242	42	7
有業人員	1	44,351	13,057	12,369	13,381	4,590	770	168	17	0	0
	2	41,521	10,119	10,399	13,092	5,977	1,373	471	89	0	0
	3	8,889	0	3,211	3,152	1,607	666	229	15	7	0
	4	2,052	0	0	979	675	273	85	39	0	0
	5	115	0	0	0	77	16	22	0	0	0
	6	20	0	0	0	0	10	10	0	0	0
	VV	3,052	203	664	851	710	355	147	81	34	7

第２表　（消費支出および十大費目）

		集計世帯数	世帯数分布（抽出率調整）	消費支出	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
総数		32,027	495,465	328,140	72,833	17,687	19,238	9,204	14,138	11,366	47,961	22,270	31,389	82,003
うち世帯人員が4人		9,944	155,850	335,438	76,362	15,345	20,214	8,885	14,452	10,987	47,894	33,442	32,269	75,588
有業人員	1	4,132	66,299	305,234	71,543	17,556	18,854	8,383	13,579	11,656	42,703	31,202	31,959	57,801
	2	4,201	64,868	347,740	78,472	12,932	20,621	8,917	14,876	10,538	52,141	39,485	33,681	76,078
	3	1,031	15,615	380,521	83,796	12,583	23,045	10,276	16,561	10,331	52,406	22,513	27,852	121,160
	4	324	4,851	399,962	85,083	18,500	22,490	11,763	14,096	10,812	51,310	4,509	32,769	148,628
	VV	256	4,216	379,882	82,134	24,296	22,238	7,843	14,238	10,011	43,521	49,453	31,206	94,942
	（特掲）一人または二人	8,333	131,168	326,256	74,970	15,269	19,728	8,647	14,221	11,103	47,371	35,298	32,810	66,839

以　上

住宅ローン返済中における家計の逼迫と消費行動に関する分析

谷口 裕明
ソニー銀行㈱　総合リスク管理部
参加カテゴリー：C

要旨

住宅ローンを抱える30代～40代の世帯を対象とし、家計の逼迫度を0%～100%で予測する確率モデルを極力シンプルに構築することを本分析のゴールとする。金融機関サイドからは把握出来ない消費行動を明らかにし、住宅ローンをとりまく環境の将来予測や、戦略策定の一助としたい。

1. はじめに

住宅ローン融資は事業性融資と異なり、融資実行後における債務者の実態把握が現実的に困難であり、多くの金融機関において債務者が支払困難な状況に陥り、デフォルトに至るプロセスをリアルタイムに把握していないのが現状である。今回、全国消費実態調査のデータを活用し、家計が逼迫している世帯に共通して見られる特徴を明らかにするとともに、住宅ローンがデフォルトに至るプロセスの一端を明らかにしたい。

ここで、借入金返済に充てる資金を実収入から工面できていない世帯を「家計が逼迫している世帯」とし、以下に記載の条件に該当する世帯と定義する。本分析のゴールは、住宅ローンの返済を行っている世帯を対象に、「家計が逼迫している世帯」に該当する確率を、極力シンプルに（少数の因子で）予測するモデルを構築することであり、具体的には、AUCが0.9以上となるロジスティック回帰モデルを構築することと定める。

前提として、住宅ローンは主に30代～40代がメイン層となるため、この世代に限定しデータを絞り込むこととし、また、30代と40代では、収入、世帯人数、子供の年齢など、その消費行動の背景が大きく異なることが想定されることから、年代を分けて分析を実施する。

「家計が逼迫している世帯」に該当する条件・・・実収入－実支出－借入金返済※　＜　0
（※）借入金返済＝土地家屋借金返済＋他の借金返済＋分割払・一括払購入借入金

キーワード：全国消費実態調査、住宅ローン、ロジスティック回帰モデル、AUC

2. ドライバー候補（指標）の選択

擬似ミクロデータの主要な項目を従属変数、家計逼迫フラグを独立変数とするノンパラメトリック検定（ウィルコクソンの順位和検定）を実施し、ドライバー候補の選別を実施する。

検定の対象となる項目は、「家計が逼迫している世帯」に該当するか否かを決定する要素である実収入、実支出、借入金等の中から以下の通り定める。

大区分	中区分	小区分(ドライバー候補)
実収入	経常収入	勤め先収入
		事業・内職収入
		本業以外の勤め先・事業・内職収入
		他の経常収入
	特別収入	受贈金
		その他
実支出	消費支出	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
	非消費支出	直接税
		社会保険料
		他の非消費支出
借入金等返済(実支出以外の支出)		土地家屋借金返済
		他の借金返済
		分割払・一括払購入借入金返済

仮にドライバーとして有効であれば、収入に該当する項目（Youto005～Youto020）は小さければ小さいほど、支出に該当する項目（Youto038～Youto180）は大きければ大きいほど「家計が逼迫している世帯」に該当する確率は高くなると容易に想像出来ることから、前者は左側1%を、後者は右側1%を有意水準とし、ドライバー候補として採用するか否かの判定を行う。

検定結果は以下に示す通り、30代において16指標が、40代において16指標がドライバー候補として残された。

支出に該当する項目で採用されなかった項目は、どれも左側1%であれば有意と判定されており、支出でありながら小さければ小さいほど「家計が逼迫している世帯」に該当する可能性が高くなるものである。直接税や社会保険料の支払いは収入との相関が高いものであり、ここでドライバーとしては使用しないが、本結果は容易に理解出来るものである。

(30代)

Variable Variable	ラベル	Two-sample Wilcoxon Statistic	Wilcoxon Statistic Standardized	P-value Wilcoxon Test (Left-sided)	P-value Wilcoxon Test (Right-sided)	P-value Wilcoxon Test (Two-sided)	判定
Youto005	勤め先収入	392844637	-74.5907713	0		0	○
Youto006	事業・内職収入	530568535	0.17989196		0.428618694	0.857237387	×
Youto011	本業以外の勤め先・事業・内職収入	530380633	-0.00962217	0.496161367		0.992322734	×
Youto012	他の経常収入	516818212.5	-7.38652611	7.53573E-14		1.50715E-13	○
Youto019	受贈金	545166453	8.062269464		3.74454E-16	7.48909E-16	×
Youto020	その他	525121072	-2.85880594	0.002126194		0.004252388	○
Youto038	食料	601585987	38.60795061		0	0	○
Youto079	住居	542145417	6.715098125		9.39699E-12	1.8794E-11	○
Youto084	光熱・水道	562904700	17.83140588		7.07034E-70	1.41407E-69	○
Youto089	家具・家事用品	598035663	36.68263903		6.9076E-295	1.3815E-294	○
Youto099	被服及び履物	595549941	35.33467331		8.6205E-274	1.7241E-273	○
Youto117	保健医療	585888018	30.09508405		2.8095E-199	5.6189E-199	○
Youto122	交通・通信	614800591	45.77412201		0	0	○
Youto129	教育	587031440	31.05101712		5.5268E-212	1.1054E-211	○
Youto133	教養娯楽	601969715	38.81604316		0	0	○
Youto142	その他の消費支出	590090000	32.37377113		3.2117E-230	6.4234E-230	○
Youto160	直接税	500652945	-16.1278443	8.12988E-59		1.62598E-58	×
Youto164	社会保険料	501517683.5	-15.6588621	1.44511E-55		2.89022E-55	×
Youto169	他の非消費支出	527430880.5	-3.10809922	0.000941474		0.001882948	×
Youto178	土地家屋借金返済	539173430	4.762132731		9.57788E-07	1.91558E-06	○
Youto179	他の借金返済	542416065.5	6.812834367		4.78471E-12	9.56943E-12	○
Youto180	分割払・一括払購入借入金返済	569877667	21.41283251		5.0728E-102	1.0146E-101	○

※この論文の内容は、全て執筆者の個人的な見解であり、所属する組織の公式的な見解を示すものではありません。

(40 代)

Analysis Variable	ラベル	Two-sample Wilcoxon Statistic	Wilcoxon Statistic Standardized	P-value, Wilcoxon Test (Left-sided)	P-value, Wilcoxon Test (Right-sided)	P-value, Wilcoxon Test (Two-sided)	判定
Youto005	勤め先収入	1492671066	-86.1926448	0		0	○
Youto006	事業・内職収入	1826631355	-13.2366646	2.69447E-40		5.38894E-40	○
Youto011	本業以外の勤め先・事業・内職収入	1830407675	-10.1634573	1.44328E-24		2.88655E-24	○
Youto012	他の経常収入	1843682529	-5.30130029	5.74904E-08		1.14981E-07	○
Youto019	受贈金	1857312025	-2.11027189	0.017417472		0.034834943	×
Youto020	その他	1801792513	-14.9034293	1.56544E-50		3.13089E-50	○
Youto038	食料	2044143581	40.98848133		0	0	○
Youto079	住居	1938520044	17.59511546		1.34254E-69	2.68508E-69	○
Youto084	光熱・水道	1940972373	17.19503728		1.44636E-66	2.89272E-66	○
Youto089	家具・家事用品	2043258899	40.78445518		0	0	○
Youto099	被服及び履物	2042021439	40.49908945		0	0	○
Youto117	保健医療	2049089239	42.12909075		0	0	○
Youto122	交通・通信	2156464381	66.89201319		0	0	○
Youto129	教育	2155091683	66.72555156		0	0	○
Youto133	教養娯楽	2050406906	42.43293542		0	0	○
Youto142	その他の消費支出	2134748681	61.88391727		0	0	○
Youto160	直接税	1786073447	-18.5288605	6.04092E-77		1.20818E-76	×
Youto164	社会保険料	1739983325	-29.1587235	3.238E-187		6.4761E-187	×
Youto169	他の非消費支出	1857233166	-3.39862199	0.000338631		0.000677263	×
Youto178	土地家屋借金返済	1847550505	-4.35000433	6.80674E-06		1.36135E-05	×
Youto179	他の借金返済	1853051843	-3.14294092	0.000836298		0.001672596	×
Youto180	分割払・一括払購入借入金返済	1999851792	30.77393609		2.9255E-208	5.851E-208	○

3. ロジスティック回帰モデルの構築

前項で選択されたドライバー候補（指標）を独立変数、「家計が逼迫している世帯」に該当するかどうかを示すフラグ（該当する場合は 1）を従属変数とし、ロジスティック回帰によるモデル構築を行う。

ここで、変数選択はステップワイズ法にて行い、変数の追加・削除ともに有意水準を 1%とする。

変数選択のステップ数は、構築モデルの AUC が初めて 0.9 を超えたとき以降、次のステップには進まないものとする。

(30 代)

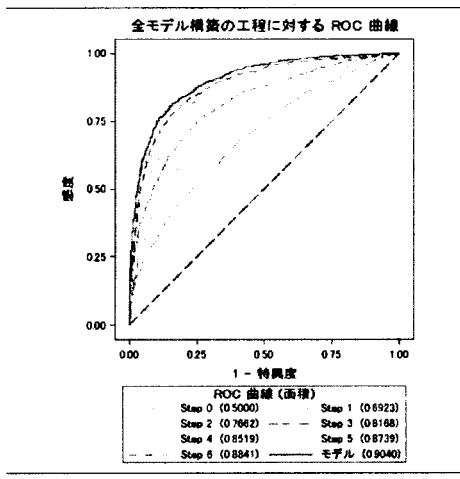
選択された変数は 7 つ。パラメータ推定値の符号と P 値は問題なく、推定相関行列の水準から多重共線性も問題なし。収入サイドの勤め先収入がステップ 1 で選択されており、相応の存在感を示すなか、支出サイドでは、教養娯楽、家具・家事用品のオッズ比が比較的高く、土地家屋借金返済のオッズ比は比較的低い。

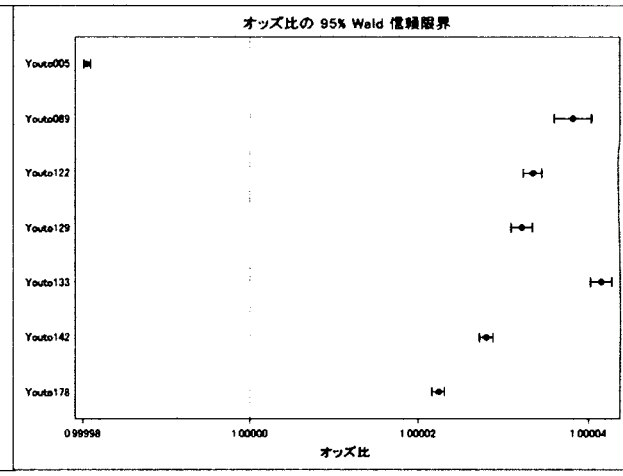
ステップワイズ法の要約								
	効果			取り込んだ数	スコア	Wald		変数
ステップ	入力済	削除済	自由度		カイ 2 乗	カイ 2 乗	Pr > ChiSq	ラベル
1	Youto005		1	1	3978.4281		<.0001	勤め先収入
2	Youto133		1	2	5736.4394		<.0001	教養娯楽
3	Youto122		1	3	5068.6254		<.0001	交通・通信
4	Youto142		1	4	4375.346		<.0001	その他の消費支出
5	Youto129		1	5	3762.3569		<.0001	教育
6	Youto089		1	6	3969.6445		<.0001	家具・家事用品
7	Youto178		1	7	3744.3479		<.0001	土地家屋借金返済

※この論文の内容は、全て執筆者の個人的な見解であり、所属する組織の公式的な見解を示すものではありません。

最尤推定値の分析					
パラメータ	自由度	推定値	標準誤差	Wald カイ2乗	Pr > ChiSq
Intercept	1	0.8329191585	0.0430000000	374.6213	<.0001
Youto005	1	-0.0000194956	0.0000001880	10759.26	<.0001
Youto089	1	0.0000381943	0.0000011000	1205.4256	<.0001
Youto122	1	0.0000335493	0.0000005453	3785.2294	<.0001
Youto129	1	0.0000322461	0.0000006435	2511.2609	<.0001
Youto133	1	0.0000415363	0.0000006350	4278.21	<.0001
Youto142	1	0.0000279867	0.0000004307	4221.7625	<.0001
Youto178	1	0.0000223091	0.0000003942	3202.4207	<.0001

推定相関行列								
パラメータ	Intercept	Youto005	Youto089	Youto122	Youto129	Youto133	Youto142	Youto178
Intercept	1	-0.5529	-0.0278	-0.0662	0.0042	0.0318	0.0427	-0.015
Youto005	-0.5529	1	-0.2175	-0.4495	-0.3383	-0.4974	-0.5236	-0.5233
Youto089	-0.0278	-0.2175	1	0.0505	0.091	0.0339	0.0979	0.1161
Youto122	-0.0662	-0.4495	0.0505	1	0.1542	0.2021	0.1287	0.2395
Youto129	0.0042	-0.3383	0.091	0.1542	1	0.1502	0.17	0.1256
Youto133	0.0318	-0.4974	0.0339	0.2021	0.1502	1	0.1727	0.2075
Youto142	0.0427	-0.5236	0.0979	0.1287	0.17	0.1727	1	0.2303
Youto178	-0.015	-0.5233	0.1161	0.2395	0.1256	0.2075	0.2303	1





(40代)

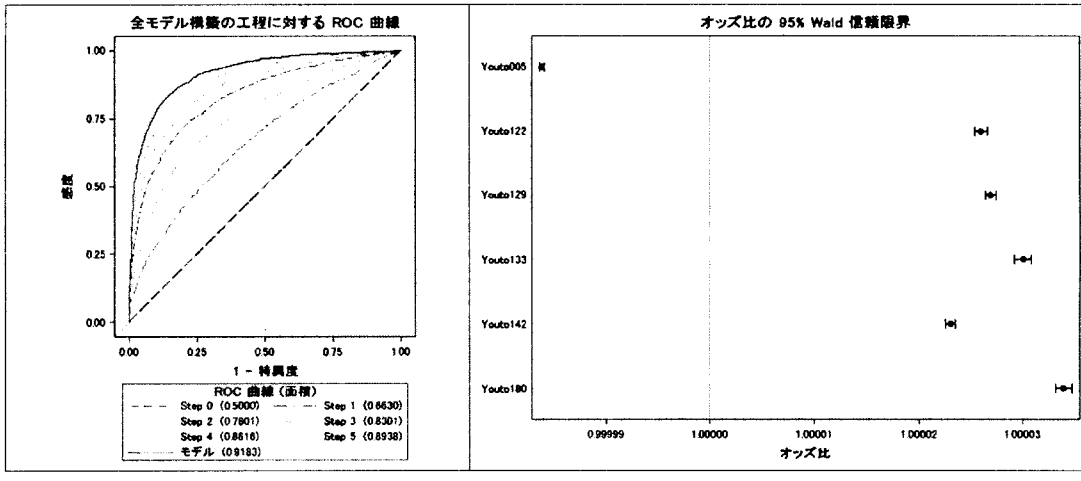
選択された変数は6つ。パラメータ推定値の符号とP値は問題なく、推定相関行列の水準から多重共線性も問題なし。収入サイドの勤め先収入がステップ1で選択されており、相応の存在感を示すなか、支出サイドでは、分割払・一括払購入借入金返済、教養娯楽のオッズ比が比較的高く、その他の消費支出のオッズ比は比較的低い。

ステップワイズ法の要約								
	効果			取り込んだ数	スコア	Wald		変数
ステップ	入力済	削除済	自由度		カイ2乗	カイ2乗	Pr > ChiSq	ラベル
1	Youto005		1	1	6136.8191		<.0001	勤め先収入
2	Youto142		1	2	13194.592		<.0001	その他の消費支出
3	Youto129		1	3	16097.265		<.0001	教育
4	Youto133		1	4	9447.7315		<.0001	教養娯楽
5	Youto122		1	5	9494.2216		<.0001	交通・通信
6	Youto180		1	6	7256.7871		<.0001	分割払・一括払購入借入金返済

※この論文の内容は、全て執筆者の個人的な見解であり、所属する組織の公式的な見解を示すものではありません。

最尤推定値の分析					
パラメータ	自由度	推定値	標準誤差	Wald カイ2乗	Pr > ChiSq
Intercept	1	1.5073880678	0.0308000000	2387.742	<.0001
Youto005	1	-0.0000162808	0.0000001152	19960.492	<.0001
Youto122	1	0.0000259915	0.0000003249	6397.8386	<.0001
Youto129	1	0.0000269216	0.0000002761	9508.0609	<.0001
Youto133	1	0.0000300534	0.0000004120	5322.1575	<.0001
Youto142	1	0.0000230351	0.0000002218	10786.1	<.0001
Youto180	1	0.0000339548	0.0000004089	6894.2566	<.0001

推定相関行列							
パラメータ	Intercept	Youto005	Youto122	Youto129	Youto133	Youto142	Youto180
Intercept	1	-0.6452	-0.0128	0.1692	0.0038	0.1815	0.204
Youto005	-0.6452	1	-0.4298	-0.5181	-0.4405	-0.6263	-0.57
Youto122	-0.0128	-0.4298	1	0.1802	0.1606	0.1672	0.1882
Youto129	0.1692	-0.5181	0.1802	1	0.2148	0.3179	0.2651
Youto133	0.0038	-0.4405	0.1606	0.2148	1	0.2322	0.129
Youto142	0.1815	-0.6263	0.1672	0.3179	0.2322	1	0.3234
Youto180	0.204	-0.57	0.1882	0.2651	0.129	0.3234	1



4. 今次構築モデルより

30代と40代の今次構築モデルを比較すると、勤め先収入、教養娯楽、交通・通信、その他の消費支出、教育といった項目が共通しているのに対し、家具・家事用品、土地家屋借金返済が30代のみに、分割払・一括払購入借入金返済が40代のみに選択されており、家計逼迫に至る背景が年代によって少なからず異なることがわかる。

まず、共通している項目を見ると、勤め先収入の水準が低ければ家計が逼迫する可能性が高いというのは大方の予想通りであるが、教養娯楽や教育といった項目は子育てに大きく関わる項目であり、多少の無理をしてでも子供への投資を惜しまないという親心が見て取れる部分である。また、交通・通信や

※この論文の内容は、全て執筆者の個人的な見解であり、所属する組織の公式的な見解を示すものではありません。

その他の消費支出は自動車や趣味など、多かれ少なかれ無駄遣いが家計を圧迫している側面があるのではないかと推察される。

次に、30代のみで選択されている項目を見ると、住宅ローンの返済が相応に家計を圧迫する中で、日常生活で必要な消耗品にかける出費をいかに減らせるかなど、家事における節約・浪費といったスタンスの差異が家具・家事用品といった項目に表れていることも十分に考えられる。

また、40代のみで選択されている分割払・一括払購入借入金返済は、30代と比べると年収の水準が高い中、家計のキャッシュフローが回らず、借入金による消費を余儀なくされている世帯が一部存在していることを示唆している。

5. おわりに

今回の分析から、大きく「子育て」と「浪費」という2つの要因が住宅ローンを抱える世帯の家計を圧迫し得ることがわかった。

前者について、少子高齢化が加速度的に進んでいる背景を如実に示すものであり、行政の早期の対策が待たれるところである。また、金融機関としても、子育てを主因とする家計の逼迫が極力回避されるよう、柔軟な対応が期待される。

後者について、現行の住宅ローンの審査では、債務者に浪費癖があるかどうかまでは十分に見ていないことが多く、融資実行後の取引振りもモニタリングされていないことが一般的であり、金融機関としては一定程度こういった層を取り込んでしまうことは不可避である。今後は、大容量のトランザクションデータを活用した浪費癖の有無の判別など、一層のリスク管理高度化が期待される。

以　上

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**
世　話　人 （氏名50音順）	伊藤 陽一	北海道大学
	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
	堺 伸也	イーピーエス株式会社
	坂巻 英一	宮城大学
	菅波 秀規	興和株式会社
	周防 節雄	兵庫県立大学
	高橋 行雄	BioStat研究所株式会社
	高部 勲	総務省統計局統計調査部経済基本構造統計課
	八木 章	近畿大学
	山之内 直樹	第一三共株式会社

協賛

株式会社ACRONET
アマゾン データ サービス ジャパン株式会社
イーピーエス株式会社
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
ファーマ・コンサルティング・グループ・ジャパン株式会社
富士通株式会社

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：03-5485-7858（開設時間：10:00～17:00）
※但し、12:00～13:00の間、また土日祝日を除く。
E-mail：sasuser2014@sascom.jp

SAS ユーザー総会 2014

医療、政府・自治体、大学による
エコシステムの実証

論文集

2014年7月24日 初版第1刷発行
発行：SASユーザー会　SAS Institute Japan株式会社