SSE-DP-2025-7

講義録：社会の中の統計科学

国友直人編

（統計数理研究所）

2025 年 10 月

SSE-DP(ディスカッションペーパー・シリーズ)は以下のサイトから無料で入手可能です。
https://stat-expert.ism.ac.jp/training/discussionpaper/

SSE-DP-2025-7

Lectures on Statistical Sciences in Japanese Society

edited by

Naoto Kunitomo

The Institute of Statistical Mathematics

October 2025

(Summary)

A lecture series on statistical Sciences in Japanese Society was delivered by several statisticians for the Statistical Expert Training Program from April to June 2025 at the Institute of Statistical Mathematics, Tokyo, Japan. The topics of lectures cover from the population problem, marketing, quality control, government statistics, financial statistics and survey sampling in Japanese society.

講義録：社会の中の統計科学*

国友直人† （編集）

2025 年 10 月

鍵言葉 (Key Words)： 社会の中の統計科学、連続講義シリーズ, 統計エキスパート人材育成事業, 統計学・統計科学の役割

要約

統計数理研究所が推進している統計エキスパート人材育成事業の一環として 2025 年 4 月～6 月にかけて「連続講義シリーズ：社会の中の統計科学」が企画された。現代の日本社会では統計学・統計科学は学術界、民間企業、政府、教育界など多様な分野において活用されているが、この傾向は今後ますます増加すると予想されている。統計学・統計科学が便利に利用されている反面、誤解・誤用・乱用される側面も存在するので、統計エキスパートはそうした面を総合的に理解しておくことが必要であり、必要が生じれば各場面で的確に対処する責任がある。統計的アプローチが役立っている具体例はきりがないほど多数に上るが、負の側面の例としては、医療統計での新薬の開発を巡る不正問題、品質管理でのデータ偽造問題、公的統計での公的データの不適切な処理問題、などの事例が近年の日本では報道されている。各講演は統計学・統計科学がよく利用されている社会における幾つかの重要な分野に絞って、統計学・統計科学が有効に利用されている側面と共に負の側面にも言及し、課題を正しく理解し、今後の統計科学の教育の現場において適切に対処する能力を養うことを目的に行われた。この連続講義シリーズは日本の各分野を代表する専門家による講演により構成されたが、敬称略で椿広計 (統計数理研究所), 金子隆一 (明治大学, 元社人研), 照井伸彦 (東京理科大学), 千野雅人 (統計数理研究所, 元統計局）, 吉田靖 (東京経済大学), 山下智志（統計数理研究所）, 鈴木督久 (日本世論調査協会会長)、の各先生にご協力いただいた。

*統計エキスパート統計エキスパート人材育成事業において 2025 年度に実施された連続講義シリーズ「社会の中の統計科学」の報告書である。なおこの報告書の内容は統計数理研究所の見解を反映するものではない。

†統計数理研究所

目次

はじめに

統計数理研究所が推進している統計エキスパート人材育成事業の一環として2025年4月〜6月にかけて「連続講義シリーズ：社会の中の統計科学」が企画された。現代の日本社会では統計学・統計科学は学術界、民間企業、政府、教育界など多様な分野において活用されているが、この傾向は今後ますます増加すると予想されている。統計エキスパート事業の関係者は様々な分野で統計エキスパートとしてこれからの活躍が期待されている。統計学・統計科学が便利に利用されている反面、誤解・誤用・乱用される側面も存在するので、統計エキスパートはそうした面を総合的に理解しておくことが必要であり、必要が生じれば各場面で的確に対処する責任がある。統計的アプローチが役立っている具体例はきりがないほど多数に上るが、負の側面の例としては、医療統計での新薬の開発を巡る不正問題、品質管理でのデータ偽造問題、公的統計での公的データの不適切な処理問題、などの事例が近年の日本では報道されている。こうした事例の中で特に医療・薬学分野にかかわる問題については、特にメンターによる「生物統計・医療統計」の講義の中でも扱われている。本講演シリーズは統計学・統計科学がよく利用されている社会における幾つかの重要な分野に絞って、統計学・統計科学が有効に利用されている側面と共に負の側面にも言及し、課題を正しく理解し、今後の統計科学の教育の現場において適切に対処する能力を養うことを目的に開講された。連続講義シリーズは日本の各分野を代表する専門家による講演により構成された統計エキスパート養成事業における重要な講義シリーズであり、2025年度に行われた講義の内容と講師は以下の通りであった。

＜2025年度の講義＞

2025.4.14(月) 「品質管理と統計科学」椿広計 (統計数理研究所)

2025.4.21(月) 「人口統計学と日本社会」金子隆一 (明治大学, 元社会保障・人口問題研究所)

2025.4.28(月) 「マーケティングの統計科学」照井伸彦 (東京理科大学)

2025.5.12(月) 「公的統計と政策・政治」千野雅人 (統計数理研究所, 元統計局長)
2025.5.16(金) 「生命保険と日本社会」吉田靖 (東京経済大学)
2025.5.26(月) 「金融リスクと統計科学」山下智志 (統計数理研究所)
2025.6.2(月) 「世論調査と統計学」鈴木督久 (日本世論調査協会会長)

この「講義録 : 社会の中の統計科学」は講演後に各講演者から「統計エキスパート人材育成事業」の研修生の為の基礎資料として提供された資料をまとめてディスカッション・ペーパー (DP) として公表するものである。ここで掲示する各資料の内容は必ずしも今年に行われた各講義と厳密に一対一に対応するのもではないが、各界の統統計エキスパート人材育成事業に参加している研修生を始め、より広く統計科学を志す関係者の勉学・教育に資するのではと判断、今後のさらなる議論の為に掲示しておくこととした。元よりここに収録した各論考は各担当講演者の責任により独立に作成され、相互に調整を加えていないが、読者からのコメントを歓迎する。

2025 年 10 月
国友直人 (連続講義シリーズ・コーディネーター, kunitomo アットマーク ism.ac.jp)

椿広計(統計数理研究所)

医学･生物統計の社会的役割／医薬品許認可制度

１． 統計科学と臨床試験

統計は、国の制度に取り込まれると同時に、学術制度にも取り込まれています。その代表的な分野が医学生物学統計分野です。1980 年代から、医学系の学術論文の審査には、必ず統計家がレフェリーに入るようになり、学術的な価値評価以外に、統計的な処理が正しいかどうかを判断します。

この講義では、こうした特別な分野である医学・生物統計の社会的役割にして話します。たとえば、新しい医薬品についての実験を論文にする場合、実験データの内容、データ解析の方法などすべてについて事前に登録します。これほど厳格な分野は、医学・薬品分野にしかありません。それ以外の学術分野では、データを採取し分析した結果、想定された結果が出なかった場合、分析方法を変更して論文にする方法もよくとられますが、医学・薬品分野では、それは許されていません。

この分野のデータには、生物や人間に対する実験が含まれるため、統計はきわめて倫理の問題と密接に絡んできます。この後話す予定の「ヘルシンキ宣言」の理念である、統計の科学性と倫理の問題が大きく関わってきます。ここでは、そうした問題を念頭におきながら話していきたいと思います。この分野では、日本の医学分野のパイオニアである佐久間昭先生がよくおっしゃるように「クスリはリスク」ということを忘れてはならないのです。

ここでは、まず統計は実験をどう考えているかについて解説します。臨床試験は、基本的には、薬品の開発の最終段階において、その薬の安定性・有効性を人間によって検証します。動物で効果があっても、人間に有効である

という保証はないので、こうした臨床実験が義務づけられています。実際、21 世紀になっても、動物に対して効果があった物質について、人間に対して安全性の実験をしたところ、８人死亡した例がイギリスで報告されています。

1.1 薬効判定に統計学がなぜ使われるのか？

この講義の冒頭部分で、薬害肝炎において、私が証人として証言した内容の一部を紹介しましたが、ここで、その続きを紹介しておきます(一部、内容を変更しない範囲で表現を修正しています)。

- 科学の諸分野に統計が応用できるということですけれども、医学や生物学の分野を例に挙げられたのはなにか理由があるのでしょうか？
  ―もちろん医学における統計学の適用・応用というのは非常に重要なことです。医学、生物学、心理学という分野は基本的に不確実性が非常に大きい。個体差、個体変動というものに支配されているということが多いわけですね。特に事実に基づく検証ということが多く要求されているということです。
- 薬効判定に統計学を応用するというのも同じ意味があるということですか？
  ―もちろんその通りです。基本的に、現在、薬効判定において事実に基づく、エビデンスに基づく検証というのは不可欠かと存じます。そういう分野における統計の専門家というものも十分確立していると考えております。
- 先ほど，統計学自体が仮説を検証する道具という側面と仮説を探索する道具となるという側面とご説明して頂きましたけれども，薬効判定の分野はどちらの側面として役に立つのでしょうか？
  ―基本的に、私のような新薬品の許認可などに関わった人間にとっては、最終的に仮説を検証するということが、大変重要なミッションと考えております。実際にいわゆる臨床試験に関わる統計的方法という部分に関しては、仮説検証型の方法論というものが多く開発されているわけです。
- これ(甲 A 第 133 号証：椿、藤田、佐藤（1998)「誰がための臨床統計」統計

数理 46 巻)は何について書かれたものでしょうか？

―これはある意味で、特に 1970 年代から日本で行われていた、日本独自の臨床試験の考え方――これは佐藤倚男先生が開発されている部分が多いんですけれども――この考え方を総括的にレビューしたものです。今非常にこの分野もグローバル化が進んでいますが、1990 年代後半以降に、いわゆる ICH といわれている国際標準的な部分と日本の考え方の間にどういう意味で共通点があり、どういう部分で相違点があるというようなことを総括しております。

なお、当時は、まだ日本では臨床試験はまともに行われていないにもかかわらず、このことについて書かれた本があります。それが、中谷宇吉郎の『科学の方法』(岩波新書、1958 年)です。彼は、雪の結晶の研究で世界的に有名な日本を代表する科学者であり、寺田寅彦の高弟としても知られています。

『科学の方法』の中では、科学の方法についていろいろ書かれていますが、科学は再現可能であることを信用することであるととらえ、その例として、薬の例をあげ、次のように述べています。

―ある人がある薬を飲んだときに、病気が治ったら、その薬は利いた、とそう簡単にいってしまうことはできない。全然同じ体質の人が二人いて、同じ病気になって、一方は飲んで治り、一方は飲まなくて治らなかったという場合でないと、薬が効いたかどうかを確かめることはできないはずである。同じ条件の人が二人いることはないから、確かめてみることができない。それで偶然に治ったのだと、あくまでも言い張られたら決め手がないのである。

―しかし、こういう場合に、科学はそれを取り扱う方法を持っている。それは統計という方法である。……実際に全く同じ条件ということはないのであるから、広い意味で言えば、科学は統計の学問ともいえるのである。

1958 年時点では、日本では統計に基づく臨床試験は始まったかどうか、という段階でした。さらに、次のような記述もあります。

―科学が統計の学問であるとすると、全ての法則には例外がある。そして，科学が進歩するということは、その例外の範囲をできるだけ縮めていくことである。

これは、すでに述べたカール・ピアソンの自由倫理の思想の発想にきわめて近いものがあります。実は、ピアソンは、統計的な実験についてはほとんど貢献していません。むしろ、次の世代である R.A.フィッシャーが統計的な実験という方法論を開発しました。医薬品の分野はそれに呼応し、医薬品の許認可は統計的な実験に基づいて行なうことが制度として確立しました。

制度化確立のきっかけとなったのは薬害事件、特にサリドマイド事件です。これは、妊娠初期の妊婦がサリドマイドを服用したことによって、世界中で奇形児が誕生したという大変悲惨な事件でした。これ以後、世界中の新医薬品審査に影響を与えることとなりました。

この中で、アメリカは唯一、サリドマイドを承認しなかった国ですが、1962 年に、自ら Kefauver-Harris 修正法を策定し、医薬品の審査を科学的に行なうよう定めました。日本は、1967 年に「医薬品の製造承認等に関する基本方針」を作成し、統計的なデータに基づく臨床試験を行なうことを義務づけました。1971 年には、「薬効問題懇談会の答申について」において、1967 年以前に申請された統計的な裏付けのない医薬品について、再見直しや再評価を行なうことを定めました。それに従って、科学的臨床試験に基づく承認、科学的証拠に基づく審査、精密かつ客観的な観察、比較試験などの原則が確立されました。

もちろんそれまでに、日本の統計家が医薬品の許認可の基本方針に関わったというわけではありません。しかしサリドマイド裁判においては、われわれの大先輩である益山吉三郎、吉村功先生などが、サリドマイドの危険性を訴えるデータについて証拠能力があることを証言されています。現在は、こういう薬害問題について国の体制は改善されていますが、近年でも、ソリブジンやイレッサなど副作用の問題はまだ発生しています。

1.2　実験概念の進化／ベーコンからフィッシャーへ

先ほどから、臨床試験は統計的実験であると何度も繰り返して説明しています。では、統計的実験以外にどんな実験があるでしょうか。実は、「やみくもな実験」というものがあります。つまり、実験したら何か結果が出てくる

だろうという錬金術的な実験です。

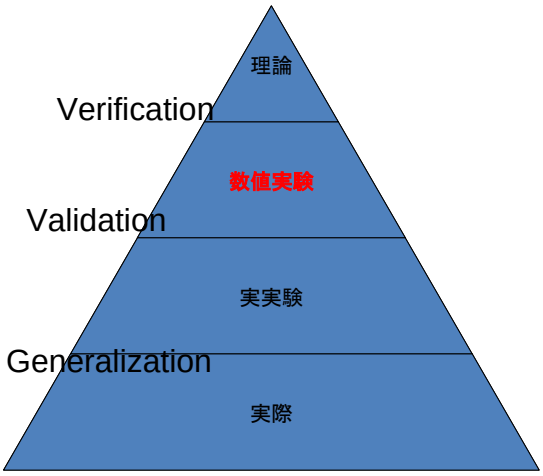
実験はある仮説を検証するための方法ですが、そのことを強調したのが、科学的精密実験です。これは今でも大学、研究機関、企業の実験室などで多く行なわれています。つまり、「できるかぎり実験条件を細かく管理し、バラツキや偶然的変動をなるべく小さくし、一つの要因の効果が明確に現われるような条件を作り出し、その因果関係を把握する」というものです。このことを最初に提唱したのはフランシス・ベーコンです。彼は、16 世紀のイギリスにおける経験主義の元祖的存在で、錬金術的な実験を批判し、上のような科学的精密実験を提唱しました。ベーコン自身は実際には実験しなかったようですが、考え方としてはピアソンの先輩にあたると言えるでしょう。文学の世界では、エッセイという文体を初めて作ったとも言われています。

ピアソンの次の世代であり、統計学の先達になるのがフィッシャーです。彼は、ロンドン郊外にあるロザムステッドの農事試験場の研究者として品種改良に取り組んでいました。農場の環境は一定ではなく、日あたり、水はけなどさまざまなバラツキがあるため、精密な生物実験はなかなか実施できませんでした。さらに、もしある農場でよい成果が得られたとしても、同一の条件で実験できないため、誤差を伴う他の農場で同じ成果が得られるかどうかに悩んでいました。

ここで、調査と実験との違いについて考えてみましょう。実験は、何かの刺激を与えるなど、何らかの介入を行ない、それに対してどのような反応があるかのデータを集めます。そして、その結果が有効なものであるかどうかを実証します。医薬品の場合は、人間にある薬を投与することが有効であるかどうかを実証ことになります。その意味では、実験は、「計測(Measurement science)」行為であることは明らかです。ただし、ここで重要なのは、個々の対象の結果ではなく、集団としての計測結果です。いずれにしても、実験と調査の大きな違いは、介入の有無でと言えます。

実験概念の進化について改めて整理すると、【図表 1】のようになります。最初は理論と実験が分離していた闇雲な実験から始まり、次にベーコンは、理論検証のための実験として、実験室内で再現性のある実験を提唱しました。これは現在でも、精密科学実験として定着しています。

【図表 1】実験概念の進化



それに対して、フィッシャーが考えた実験は、実際に使える結論を出すための実験でした。つまり、ある薬が有効であるという仮説を検証するための実験などです。そして現場では実験のノイズはコントロールできないので、ノイズの許容と確率変数化を試みました。同時に、実験結果を社会に対して一般化できるための実験方法も考えました。つまり理論検証のための実験ではなく、実際に使える結論を導き出すための実験です。現在、医薬品の臨床試験で行なわれている実験は、フィッシャーの方法論に基づいています。

２. 新医薬品候補物質の開発と臨床試験

2.1　臨床試験の3つの相

医薬品の候補物質を発見するために、どんな臨床試験が行なわれているかを紹介します。有効と思われる候補物質が見つかると、まず試験管の中で実験し、次に、動物でその効果が再現できるかどうか、薬理試験・毒性試験・吸収・排泄試験などの一連の前臨床試験を行ないます。その上で、1962 年からアメリカが薬事法を改正して採用した臨床試験のフェイズシステムに基づいて、臨床試験を行ないます。臨床試験は３つの相によって行なわれます。

◆第１相（安全性）

男性健常者だけを対象にして安全性の確認をします。

◆第２相（有効性）

患者を対象にして、治療における用法・用量の確認、用量反応関係などを調べます。

◆第３相（大規模試験による検証）

患者を対象にして、プラセボないしは標準治療を対照とした二重盲検比較試験を行ないます。二重盲検承認については、後ほど詳しく説明します。

医薬品は、これらの相ごとの臨床試験をクリアして初めて承認されます。市販後も必要に応じて第４相試験が行なわれることもあります。たとえば、実際の治療の現場における安全性に関する情報収集調査、真のエンドポイントにおける有効性の比較試験などです。

ここで、「真のエンドポイント」について補足しておきます。医薬品を国が承認する場合、サロゲート・エンドポイント（代替的エンドポイント）で承認する場合もあります。たとえば、高脂血症に対して、コレステロールを下げる薬があります。実は、コレステロールが高いこと自体は病気かどうかはかなり疑問で、コレステロールが高いことによって、心疾患や脳血管系の病気がおこりやすいだけです。高脂血症という病気自体、コレステロールを下げる薬ができたことによって生まれた病気でもあるのです。

1990 年代に、スタチン系のメバロチンという、コレステロールを画期的に下げる薬が日本で初めて承認されたことによって、コレステロールを下げることが治療目的になりました。当時の医薬界の常識では、コレステロールを下げれば、虚血性心疾患などが予防できると期待しました。そこで、スタチン系の薬を飲む習慣が世界的に広がりました。しかし問題は、メバロチンが承認されたとき、まだ心疾患を予防する効果があるかどうかは分かっていなかったことです。

本当は、メバロチンに心疾患を予防する効果があるかどうか、10 年、20 年と長期にわたって調査をした後、承認すべきなのですが、コレステロールを低下させる効果が分かった段階で承認したほうが、社会に対するベネフィットは大きいと判断されたわけです。つまり、真のエンドポイントは心疾患の予防ですが、代替的なエンドポイントとしてコレステロール低下薬を認可し

たことになります。その後、メバロチンを開発した三共製薬に市販後調査を義務づけ、虚血性心疾患を予防するかどうか8000例で比較試験を行ないました。これは、当時の日本で最大の臨床実験でした。

日本の臨床試験がしばしば海外から批判されたのは、個々の計測行為の質が低いのではないかという点です。昔は、血圧やコレステロールなどの計測を個々の医師に任せていたため、誤差が大きいのではないかと批判されました。また計測行為としてのエンドポイントについて、日本は、薬が患者に効いたかどうかについて、客観的な測定値より、医師の主観的判断を重視しました。1990 年代末までそういう状況が続いたので、それに対して、特にアメリカから批判されていました。

2.2　日本の臨床試験の転換点

日本の医薬品の臨床試験は、1967 年に非常に大きな転換点がありました。このときから科学的方針に基づいて臨床試験を行なうことが、国の方針として定められました。それまでは、日本の医薬品の承認方法は、俗に「三た論法」と言われていました。たとえば、医師が「ある薬を私の患者さん20名に使ったところ、大半の方で症状が改善しました。従ってこの薬は有効です」と言えばそれが根拠になるわけです。つまり、「使った、効いた、良かった」の「三た論法」で、実際そういう報告書で承認された例はたくさんあります。

当時はそれで仕方なかったという状況も理解できないわけではありません。薬害訴訟で問題になったフィブルノゲン製剤の承認もこの時期のことで、その承認理由も「60 人くらいの患者に使ったら全員に効果があった」という３行程度の文章があるだけです。どういう症状の患者にどのように使ったかという記録はいっさいありません。当時はそれほど異例の状況ではなく、簡単な報告での承認が一般的だったのです。

「三た論法」以外では、「雨乞いの太鼓」という考え方もありました。雨乞いの太鼓は必ず雨を降らせられるという神話があるのですが、実際は雨が降るまで太鼓をたたくからなのです。同様に、治るまで薬を使えば、必ずその薬は効くわけです。この場合、使わなくても治るのではないか？という疑問

に対しては、先の中谷宇吉郎の指摘のように、使った場合と使わなかった場合の比較試験をしなければ結論は出せません。「三た論法」も「雨乞いの太鼓」神話も今でもまかりとおっていますが、医薬品のような厳しい分野では、こういうあいまいな根拠は通用しません。

2.3　壊血病と脚気をめぐる歴史的事実

この分野でよく例としてあげられるのが、レモンと壊血病についてです。東インド会社が 1600 年に４隻の船でインド遠征に出発しましたが、ほとんどの船で壊血病の患者が多数発生しました。1700 年頃でも、世界一周すると４分の３の乗組員が壊血病になるという状況だったのです。ところが、ランカスター将軍の船にだけ壊血病の患者が出ませんでした。彼らだけ、レモン汁を飲んでビタミンＣを摂取していたからだと思われます。

しかし、それについての比較試験が行なわれたのはかなり後のことで、1747 年になってからでした。1 年間の航海の船医だったリンド(Lind)は壊血病の予防のために、12 名に対して、硫酸液、リンゴ酒、酢、レモン＋オレンジ、食塩水、ニンニク＋カラシを摂取させる比較試験を行なったところ、レモン＋オレンジを摂取させた船員に効果がありました。このデータはそれなりに評価されたようですが、英国海軍がレモンを採用したのは 1795 年になってからでした。

日本においても、脚気について全く同じような歴史がありました。明治時代、当時の海軍軍医だった高木兼寛と陸軍軍医だった森林太郎(鴎外)との間で大論争がありました。海軍は、1883 年にハワイ航海に出ますが、白米食を食べていた 376 名のうち半数が脚気になり 23 名死亡しました。高木は脚気が軍隊に多く発症するのは、白米を食べているのが原因ではないかと考えました。脚気患者が出ていない陸軍のある部隊では、少し麦を食べさせているという情報が伝わったため、それに注目して、翌年は船員には非常に嫌がられたようですが、半麦食あるいはパンに強制変更して同じ航海をしました。その結果、脚気患者は皆無でした。その後、海軍は白米食だけをやめ、麦やパンを取り入れるようにしました。

森鴎外はドイツに留学して先端的な医学を学んでいましたが、高木兼寛はイギリスで医学を学んでいました。ドイツ医学では、脚気は細菌が原因であるという説が有力でした。森鴎外は高木の方法に対して、大変鋭い指摘をします。高木は食事が原因であると主張しているが、それを実証するためには、同時に一つの船で、白米食と小麦食のグループに分けて比較しなければいけない。しかし、そういう実証実験ではなかったわけです。高木の用いた手法は、今日的には、ヒストリカル・コントロールと呼ばれるものでした。そのため、最初の船に脚気菌が存在し、後の船にはたまたま存在しなかったという批判に反論できないのではないかと主張し、実験しなおすことを求めました。

しかし高木は状況証拠から、脚気の予防には圧倒的な効果があると主張し、海軍は小麦食を採用しましたが、陸軍はその後も白米食を続けました。その後、日本は日清戦争に突入します。この戦争での戦死者は約５万人ですが、戦闘による戦死はごくわずかで、脚気での死亡が約３万人という悲惨な結果となりました。その後、陸軍も脚気予防のために食事メニューを変更しました。高木の方法はヒストリカル・コントロールとして批判されましたが、それでも時期を変えた比較によって結論を出したことが非常に重要です。

レモンと壊血病、脚気の問題が統計的に興味深いのは、いずれの時期もビタミンはまだ発見されていなかったということです。つまり、真に科学的な原因は解明されていなかったわけです。しかし、事実の観察から正しい行動を導き出すことができました。これは非常に重要なことです。ふつうは統計的な手法は原理や仮説を検証するために用いられますが、原因としての根本的なメカニズムは未知であっても、正しい行動を設計できる可能性がある方法論でもあるということです。

こうした事例は他にもいろいろあります。18 世紀の初頭、スノー(Snow)という人が、「この井戸の水を飲むとコレラになる」と警告を発しました。もちろん当時はまだコレラ菌は発見させていません。しかし、ある井戸の水を飲むとコレラにかかるという事実があれば、それに対する処置をとることはできます。つまり、統計学は社会的に未知の領域に対応できる学問であることの証左です。

2.4　比較試験その後

医薬品分野は、人間対象の比較試験をする点できわめて特殊です。人間は動物と異なり言葉を話しますし、また心理的影響も強く受けます。薬についても「効くと思えば効く」という現象がしばしば見られます。

これは余談ですが、メスメルはモーツァルトと同時代のウィーンの名医で、何でも治してしまうという評判をとっていました。彼は、磁石を使って病気を治すとされていましたが、心理効果で治っている場合も多かったようです。そのため今日、英和辞典で“メスメリズム(mesmerism)”は「催眠術」と訳されています。ことほどさように、人間は心理的影響を受けやすい面があるのです。

そこで、1933 年にはすでにエバンス(Evans)が、狭心症治療の実験でプラセボとの比較試験を行なっています。これは、フィッシャーが「実験計画法」を確立する 1935 年より前のことです。ところが患者は、自分の飲んでいる薬が本当の治療薬でなくプラセボであることを知っていれば、「使った、効いた、良かった」という効果はまったくあらわれないわけです。そこで、1937 年には早くも、患者に薬の種別あるいはプラセボ投与を認知させない方法（ブラインド法）が推奨されます(今日では、ブラインド法ではなく、マスキング法と呼ばれています)。

さらに、その後 1950 年には、患者だけではなく、治療にあたる医師にも、どちらの薬が使われているか分からないようにするダブルブラインド法(二重盲検)が実施され、現在では、このスタイルが人間を対象とした比較試験の制度として義務づけられています。

科学的に仮説を検証するためには、闇雲なことをするのは論外で、仮説を立てて実験に臨まなければなりません。医薬品の有効性・安全性・有用性に関する仮説については、たとえば「有効」を厳密に定義する必要があります。たとえば、……において、……に対して、……を行うと、……と比べて有効であることについて、すべてのプロセスを厳密に科学的に定義した上で実験をしていきます。決して、高血圧の患者と不整脈の患者を一緒にしてはいけ

ません。両者には同じ薬が効くと言われていますが、一緒に同じ実験をしてはいけないわけです。

なお、日本の臨床試験においては、プラグマティック・トライアルと言われるように、医師の裁量権はある程度認められています。私自身は日本の方法は決して否定していませんが、海外では、どんな患者にも仮説に基づいて一定の治療を行なうという原則がきわめて強く貫かれています。

また、この分野では、統計科学的に立証することを義務づけており、効くはずだから効くという原理に基づく推論は仮説にすぎないという立場をとっています。そこで統計科学によって事実に基づいて推論し、仮説を検証していきます。また、統計的実験によって証明されたもの以外は、因果関係の有効性を認めないという厳しい立場をとっています。したがって、患者の希望や医師の裁量に任せて行動するという記述的な方法論は許されていません。患者の希望や医師の裁量に任せると、効く、効かないの判断に患者や医師の判断が混ざってしまうからです。

2.5　無作為二重盲検比較試験の意義について

現在、世界の標準となっている無作為二重盲検比較試験(Well-Controlled Clinical Trials)で求められている臨床試験のあるべき姿（要求品質）は、次のようにまとめることができます。

◆ 妥当性（目的との合致)
- 目的・対象集団（母集団）の明確化
- 目的を実現するための方法・プロセスの明確化
- 評価方法の明確化

◆ 倫理的許容性
- 社会にとって最も良い行動（Virtue)とモラル人格の尊重

◆ 科学性
- 偏りの排除
- 精度の高い報告結果

◆ 結果の一般化可能性

こうしたいくつかのチェックポイントを設定し、臨床試験をすべきかどうか、していいかどうかを決定します。しかも、これらの要求を実現するプランについて、すべて事前に文書化しておかなければなりません。これが、試験計画書(プロトコル)の作成の義務化です。場合によっては、試験を実施する主体以外の機関に登録しておく必要もあります。登録されていない試験はまともな試験と認められませんし、学会誌によっては投稿も受け付けられないものもあります。

ここで、科学的実験の必要性についてもう一度整理しておきましょう。

◆比較による実証

治療した場合としない場合（プラセボ）の比較をすることによって、「今までの治療方法でも治ったのではないのか？」という批判に反論することができます。日本では、新しい治療と標準治療の比較をすることがしばしば行なわれました。アメリカでは、新しい治療とプラセボの比較が多い傾向があります。

◆対照群（Control Group)の導入

実験方法として世界中でよく行なわれているのは、たとえば 100 人を 2 つの群に分け、片方は無治療、片方は標準治療にして結果を比較する方法で「併行群間比較」と言います。逆に、現在もう行なわれていないのは、先の脚気の実験などのような「歴史対照（ヒストリカル・コントロール)」です。

なお、自分自身を対照コントロール群にすることは許されています。たとえば、目薬の場合、右目に標準薬、左目に新治療薬を使って比較する方法です。あるいは、第１期に新薬、第２期にプラセボを使うというやり方はヒストリカル・コントロールですが、自分自身に対して、最初にどちらを使うかはランダムに決めるという方法を「クロスオーバー試験」と呼んでいます。これによって、時期によるバイアスを評価することができます。このようにこれまでも自分自身を比較対照コントロール群にする方法が使われたことはありますし、今でも許されています。

◆統計科学的実験

統計科学的実験においては、無作為割付(Random Assignment)という方法を用い、新薬か標準薬かプラセボかなどをランダムに決めていきます。この患者にはこの治療法がよいのではないかという医師の裁量は認めません。どちらの案かは完全に無作為に決めていきます。当該治療をしなかった場合とした場合との比較を集団で実現していくわけです。

フィッシャーが「実験計画法」を考案した際、A薬かB薬かをくじ引きで決めるという方法に対しては、社会的に非常に大きな反発を受けました。しかし、患者の個体差が大きく、どの治療法が適切かが分からない状況では、無作為に決めることによって、患者の個体差や系統的な変動が２つの群に均等に分配されると考えられます(これについては、後に詳しく述べます)。

◆比較試験を行ってよい条件

そもそも比較試験を行なってよいかどうかの倫理的配慮が必要です。私は冒頭で紹介したように、フィブリノゲンをめぐる薬害肝炎訴訟において証言しましたが、その際、国側から反対尋問も受けました。血液製剤に関しては比較試験なしで承認していましたが、1960 年代当時は、歴史的には当然のことであり、それ自体を否定する気はありません。

一方、フィブリノゲンのような血液製剤でプラセボ対照をしなければ認可されないかどうかですが、止血効果についてプラセボを使うことが許されるかどうかという問題があります。当時はプラセボ対照は認められなかったという歴史的な状況は認めるとしても、止血に効果があるものをベースラインの治療として、そのプラスアルファとして、血液製剤を使う場合、使わない場合の最低限の試験は計画されるべきだったと考えます。

では、なぜ二重盲験が求められるようになったのでしょうか。1950 年代に、「味の素を摂取すると頭が良くなる」という論文が発表されました。これは、子どもを対象に、シングル・ブラインドで検証されたものでした。その後、数年にわたってこの真偽をめぐって論争が続きました。たしかにシングル・ブラインドではかなりきれいな成果が出ましたが、その後二重盲検で再実験すると、まったくその効果は再現しませんでした。研究者側に、味の素の効能についての思い込みがあり、その結果、子どもの活動性が高まるなどとい

う研究成果を出してしまったのだと思います。

ただ、私自身、味の素で二重盲験ができるかどうか疑問に感じるところはあります。というのも、二重盲検の場合、識別不可能性が絶対条件になります。実薬とプラセボでは、見た目、匂い、味などまったく同じにします。その点、味の素はどのようにしたのか、歴史的なエピソードになってしまいましたが、興味のあるところです。日本では、漢方薬のプラセボも作られたことがありますが、匂いも味も同じなら、それを飲むだけで効くような気もしますね。

それはともかく、プラセボ対照がどんな場合に可能かを考えてみると、私自身は、治療薬が確立している治療に対して使う気はまったくありません。日本ではそういう立場をとっています。しかしアメリカでは、治療法が確立している分野でもプラセボ対照が使われています。これは倫理的には非常に大きな問題で、その意味では、日本のほうが健全だと思います。

2.6　無作為化の効用

先ほどから無作為化について説明していますが、実は、これは非常に説明が難しい概念です。統計について少し学んだ人は、回帰分析の誤差は正規分布に従うと教わります。そして、誤差について確率分布を想定するために、統計はさまざまな分野に適応できると考えられています。

しかし、ある統計モデルを考えた場合、本当に誤差は確率分布に従ってくれるかどうか。本当のところ、統計家は、誤差は確率分布に従うという想定は社会では成立しているとは思っていません。一方、フィッシャーは、系統的な誤差がそれほどないデータを意図的に作る方法を考えました。

たとえば私がＡとＢの薬を与えられたとき、私自身は、普通の人の平均的な反応とは異なる反応をするかもしれません。しかし、Ａという薬の場合マイナス１、Ｂという薬の場合プラス２の反応だとすると、それ自体は私にはいつでも再現できる反応ですが、どちらの薬にするかをくじ引きで決めれば、どちらのグループに行くかは５割の確率となります。それを私だけではなく、実験に参加する被験者全員に当てはめると、Ａ、Ｂの薬の効果が系統的な変

動ではなく、ランダムな変動、つまり確率分布として扱える変動になってきます。同時に、私自身が治療を受けたときの環境の効果などもランダムにあちこちに当てはめられることになります。

それによって、系統的な誤差ではなく、ランダムな誤差になります。実験順序をランダムにすることによって、誤差やノイズの影響は、サイコロを振って決められたときの変数のように、確率分布に従う変数と同様にふるまうようになります。これが、わざとデタラメさを実験環境の中に入れてデザインする意味で、無作為化の理論的な根拠です。つまり、フィッシャーは、統計的な推論や推測が前提としている確率分布の想定が成立するような環境を実験の中に導入したわけです。

こうした方法に対して、最初は大きな抵抗がありました。日あたりのいいところ、日あたりの悪いところでは育ち方が違うのだから、ランダムに選ぶのではなく系統的に割り当てたほうがいいのではないかなど、さまざまな議論が行なわれました。今日、無作為化はさまざまな分野で使われ、実験だけではなく調査においても標準的な手法となっています。そこで、無作為化という手法を使えば、どんなことが可能になるのか、いくつかの例題で考えてみましょう。

〈例題１〉
最小メモリ 1㎜ の物差しで、1μm の精度でものの長さを計るにはどうすればよいか？

〈答え〉

ふつうわれわれは、物差しをきちんと当てて、1mmより小さい単位は直感的に判断します。名人では、0.1mm程度まで正確に測ることができるかもしれません。しかし、ランダム化を利用すると、さらに小さな単位まで正確に測ることができます。まず、できるだけランダムに定規を計測対象に当てます。あとは、測定を繰り返し平均をとればいいわけです。100 回、1000 回、１万回、10 万回……と繰り返していけば、精度を上げることができます。これは実は対数の法則です。

非常にバカバカしい方法のように思われるかもしれませんが、実は先端的

な計測の方法でもあるのです。たとえば水平線を超えて対象の向こうに何があるかを計測する場合、乱数によりでたらめな信号を送り、反射で返ってきたものを計測する方法がとられています。

〈例題２〉
２つの物体の重さを１回の測定で偏りなく計測するにはどうすればよいか。

〈答え〉
まずランダムに物体を選び、重さを測定します。測定した方の物体は、その重さを２倍に報告し、測定しなかった方の物体は、重さを０と報告します。きわめて現実味のない、バカバカしい方法ですが、数学的には正しい方法です。つまり、２分の１の確率で重さが２倍になり、２分の１の確率で重さが０になりますので、期待値としてはそれで偏りがないことになります。

これは標本調査の基本原理です。たとえば民主党の支持率調査をする場合、選挙区の中の有権者から 100 分の 1 の確率で無作為抽出をしたとします。有権者が 1 万人とすれば 100 人抽出されることになります。そのうち民主党に投票した人を 100 倍、しなかった人を０とカウントします。両者を合計すると、１万人の集団における民主党支持者数の偏りのない推定値を計算することができます。この考え方は、支持率の推計だけではなく、所得、売り上げ、その他さまざまな分野に適用することができます。

無作為化試験のもう一つの意義は、１回の試験で２つの治療法の有効性を同時に計ることができる点です。当たった臨床試験の結果は２倍、当たらなかった結果は０と考えるわけです。もっとも０と考えるのはあまりに非現実的なので、普通は、 他の集団の平均値に合わせるという方法をとります。いずれにしても、無作為化試験を使えば、個人レベルでは無理ですが、集計レベル・集団レベルでは確率的複製が可能になっています。

このように、今日、無作為化試験はしばしば行なわれていますが、一方、その対抗策として、患者背景を強制的にバランスさせるように割付を行う方法もいまだに行なわれています。この方法は「無作為化」に対して「最小化」と呼ばれています。

2.7　日本の歩みと新旧ガイドライン

参考までに、日本の歩みを簡単に紹介しておきます。

・1943 年　薬事法, 1960 年改正
　1955 年　ペニシリンショック死発生
　1958 年　中谷「科学の方法」で統計的臨床試験の概念解説
・精神医学会の二重盲検法採用 1956 年
　グルタミン酸論争(1946-1951)
　1957 年 GABA（ガンマアミノ酪酸）の有効性判定で二重盲検による評価実施
・1962 年　サリドマイド事件（アメリカでキーフォーバハリス修正薬事法)
　1963 年中央薬事審議会「医薬品安全対策特別部会」設置
・1962 年　高橋の大衆薬批判
　1964 年　ヘルシンキ宣言
・1965 年　中央薬事審議会に二重盲検法支持者 2 名参画（二重盲検実証データの要求開始)
・1967 年　厚生省医薬品の製造承認に関する基本方針
　1970 年　キノホルム事件
・1971 年　薬効問題懇談会「医薬品の再検討に関する答申」
　厚生省医薬品再評価の実施通達
・1979 年　降圧薬ガイドライン（以後，臨床試験のガイドライン発行続く）
・1985 年　GCP 案骨格固まる，1990 年実施
・1992 年　臨床試験の統計解析に関するガイドライン通達

日本の臨床試験は歴史的には、単純な、というか、誰でもできる臨床試験を採用し、その意味では、患者や現場の医師の立場に立ったものだったと言えます。最終的には市販薬となるので、どんな医師でも参加できるような試験というシンプルさがベースとなっていたと思われます。臨床試験の統計解析に関する旧ガイドラインでもその流れがありましたが、各国からかなり批判を受けて、グローバルスタンダードである新ガイドラインでは、科学的精密性を求める方向に振り子が揺れています。

旧統計ガイドライン時代に行なわれていた日本の臨床試験には、次のような問題がありました。

◆多重性

臨床試験の比較試験の結果、網羅的な検査項目をあげて、そのうち都合のいい結果だけを“つまみ食い”することです。仮説をたくさん用意し、多くの評価項目を並べていく手法が認められていました。仮説の多重性の問題については、すでに 1977 年の段階で、その危険性が指摘されています。

◆同等性

日本では、標準治療と最新治療を比較するために、標準薬がしばしば用いられてきました。いってみれば、ベストな治療を行なっていたわけですが、そのため、標準薬と比較して有意に劣っていなければ同等の薬であるとして承認してもいいのではないかとされました。これは明らかに統計の検定の誤った使い方ですが、国の制度の中でも、同等性が認められていたことは大きな問題だと思います。

◆不完全症例の問題

患者が途中で脱落した場合、解析の除外を極小化するという方法がとられていました。

これらの点は新ガイドラインでも継承されています。一方、グローバルスタンダードに従い、できるだけ現場に近い環境で実践的に試験を行なう指針は不採用となりました。また、有効性・安全性を総合的に判断して有用性を判断すべしとの指針も不採用となりました。

３. 医薬品分野おける調査という方法について

3.1　レンツ報告がもたらしたもの

これまで医薬品関係の臨床試験について話してきましたが、ここで調査についてもふれておきたいと思います。

ここで紹介するレンツ報告は、最初の薬害事件であり、今日の医薬品制度

をつくるきっかけとなったサリドマイド事件を調査した報告書です。妊娠初期の女性が服用すると、奇形児が生まれるリスクがあるとして、大変大きな社会問題となりました。すでに市販された薬の場合は実験が行なわれることはほとんどなく、調査で証拠が出されますが、レンツ報告も同様で、実験ではなく調査に基づいて出されたものです。

これまで説明してきたように、医薬品の許認可には統計的実験が不可欠であり、EBM(Evidence Based Medicine)においては、調査は実験よりは証拠力は小さいとされています。しかし、調査はむしろ医薬品が承認された後の副作用や安全性を検証することによって、医薬品の承認を取り消すためには非常に大きな役割を果たします。

サリドマイド事件におけるレンツ報告は、後から述べる「後ろ向き調査(Retrospective)」によって行なわれました。日本ではレンツ報告の証拠力をめぐって裁判で争われました。統計家が裁判に出廷することはほとんどありませんが、これはもっとも古い事例です。大阪大学の統計学の教授はレンツ報告の証拠力を否定しましたが、増山元三郎、吉村功両先生は、レンツ報告は統計的に妥当であると証言し、裁判所もレンツ報告の証拠能力を認めました。

3.2　関連性調査の3つの形式

調査の主な目的は第 1 講で説明したように、国勢調査など集団の特性を記述することにあります。それに対して、医薬品の分野における調査の目的は、関連性調査です。医薬品に限らず、公衆衛生、疫学などにおいては、特性間の関係性を推論する調査が行なわれます。その結果、エビデンスとして完全な情報ではないにしても、たとえば医薬品が副作用の原因になっているという因果関係に関する情報も示唆可能になります。

関連性調査の形式は、以下の 3 つがあります。

①断面的調査(Cross Sectional)
②前向き調査(Prospective　または　Follow-up)
③後ろ向き調査(Retrospective または　Case-Control)

それぞれの調査の違いを明確にするために、次のような分割表を考えてみましょう。このうち、原因系は、ある薬を「投与されている／投与されていない」、結果系は「副作用が生じた／生じていない」というふうに考えることができます。これに対して、３つの調査のパターンがあるわけです。

表１　２×２分割表の基本形式

	結果系（応答）		
原因系	B	Not B	合計
A	n_{11}	n_{12}	$n_{1\cdot}$
Not A	n_{21}	n_{22}	$n_{2\cdot}$
合計	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

■横断的にとられたデータから得られた分割表

データをとる前には、データを採る期間かデータの全個数しか定めず、得られたデータの原因特性、結果特性を調べることによって得られる頻度表を「横断的研究（Cross Sectional Study)」で得られた分割表と呼びます。横断的調査データの特徴は、頻度表の行和ni・も列和ｎ・ｊも調査開始時には決められていないことにあります。

つまり、薬を飲んだ人／飲まない人、副作用が発生している人／発生していない人を何人とるかということなどを事前には決めません。世の中全体がどうなっているかを知るには適した方法で、公的調査、官庁統計などの多くは、この方法で行なわれています。

■前向きにとられたデータから得られた分割表

データを採る前に、原因特性がＡである対象ｔ１例、Ａでない対象ｔ２例をランダムに選び、それらの結果特性がどうなるか、あるいはどうなっているかを調べることによって得られる頻度表を「前向き研究(Prospective Study)」で得られた分割表と呼ぶことがあります。この場合頻度表の行和 ni・が事前に固定されていることに注意してください。実験や追跡調査で得られる頻度

表が代表的な例です。

たとえば、タバコを吸っている人 1000 人、吸っていない人 1000 人を事前に決めて、それぞれの人を追跡調査して、そのうち肺がんになった人が、喫煙群と非喫煙群でどのくらいの割合で生じているかを調査します。こういう調査方法によれば、タバコを吸ったときに肺がんになる確率、タバコを吸わないときに肺がんになる確率を推定することができます。

■後向きにとられたデータから得られた分割表

データを採る前に、結果特性がＢである対象ｓ１例、Ｂでない対象ｓ２例をランダムに選び、それらの原因特性がどうなっているかを調べることによって得られる頻度表を「後向き研究（Retrospective Study)」で得られた分割表と呼ぶことがあります。この場合、頻度表の列和ｎ・ｊが事前に固定されていることに注意してください。後ろ向き研究は、「Case-Control Study（症例・対照研究)」とも呼ばれます。

たとえば、肺がんになっている人 1000 人、なっていない人 1000 人を選び、肺がんの原因としてタバコを吸っているかどうかを調べます。これによって、肺がんになっている人の中でタバコを吸っている人の割合、肺がんになっていない人の中でタバコを吸っている人の割合を調べることができます。

サリドマイド事件のきっかけになったレンツ報告は、この後ろ向き調査で得られた結果です。つまり、奇形児の中で母親のサリドマイド服用が何人かを調べ、奇形児ではない子どもの中に母親のサリドマイド服用が何人いたかを調べ、両者を比較したものです。この点が裁判で争点となりました。大阪大学の教授は、レンツ報告は後ろ向き研究である点を批判しました。サリドマイドを服用したときに奇形児が発生する確率と服用しないときに奇形児が発生かる確率を推定しなければいけないのだから、そのためには横断調査と前向き調査でなければならないと主張しました。

しかし実は、今日に至るまで、疫学分野や公衆衛生分野では、後ろ向き調査でリスク要因を調べるのが普通です。喫煙は健康に悪いのはすでに常識になっていますが、タバコと肺がんの相関関係について、前向き調査でエビデンスが報告されている例はきわめて稀です。一説によれば、アメリカの刑務

所で同意を得た上で、囚人を喫煙群と非喫煙群に分けて調査した例はあるようですが、大半は後ろ向き調査での報告ばかりです。

では、なぜ後ろ向き調査が使われるかと言えば、副作用やがんのように小さい発生率の要因検討をデータ数を大きくせずに行うことができるからです。こうした小さい発生率は、前向き調査ではほとんど捕捉できません。もし喫煙者に３～４割も肺がんが発生していれば、とっくにタバコは禁止されているはずです。逆に、現在不幸にして肺がんになっている人を捕捉することは可能です。率直に言えば、統計的にはそういう人を対象にした後ろ向き調査のほうがコスト的にも合理的です。

したがって、副作用やがんなど小さい発生率の中で、データ数をそれほど大きくしないで因果関係を検証する場合には、後ろ向き調査が行なわれているわけです。しかし、それでは、薬を服用したときの副作用発生率の推定は不可能です。あくまでもがんになった人の中で、ある薬を服用していた人の確率が分かるだけです。つまり、後向き研究で得られた分割表から推定可能なのは、あくまで結果特性が与えられたときに、原因特性を持っている比率(条件つき確率)なのです。

3.3　リスク比とオッズ比

サリドマイド裁判は、副作用発生率はデータからは推定できないと判断され、その点で、大阪大学の教授の主張は正しかったわけです。しかし、薬を服用した場合と服用しない場合の副作用の発生率が等しいという仮説の検定だけは後向き調査もできる、というのが統計的な正しい主張です。したがって、後ろ向き調査で得られたデータであることを忘れて、あたかも前向き調査で得られたかのように処理してよいと判断されました。このことが、副作用のみならず、疫学などでも積極的に利用され、肺がんと喫煙など、種々の疾病と習慣の関連を示唆できるようになりました。

なぜ、このように判断できるのか。それはリスク比とオッズ比の問題に関わってきます。サリドマイドを服用した際のリスク比は、サリドマイド服用者(P_A)がサリドマイド非服用者(P_B)にどのくらいの割合を占めているかで

あらわすことができます。副作用に限らず、たとえば単身者世帯とそうでない世帯で自殺発生率がどの程度違うかなどを調べるときにも使われます。

リスク比：$P_A／P_B$

リスク比自体は、前向き調査でなければ正確に把握できませんが、もう一つの概念としてオッズ比があります。オッズは、競馬などの賭けごとの際によく用いられる概念ですが、自分が勝つ確率と負ける確率の比を意味しています。オッズ比は、次の式であらわすことができます。サリドマイドを服用して奇形児が発生する確率と発生しない確率、サリドマイドを服用しないで奇形児が発生する場合と発生しない場合の確率の比を示しています。

オッズ：$O_A＝P_A／（1－P_A）$，　$O_B＝P_B／（1－P_B）$
オッズ比：$O_A／O_B$

これによって、サリドマイドの副作用が発生するリスク比をオッズ比で代替することができます。リスク比＝1 という仮説は、 薬を飲んでも飲まなくても副作用の発生率は同じことをあらわします。オッズ比＝1 であるという仮説は、実は副作用の発生率が同じであるという仮説と同じです。$P_A＝P_B$ なら$O_A=O_B$ですから、リスク比とオッズ比が同じであることは統計的には区別できません。

ここから先は厳密にはきちんと証明しなければいけないのですが、断面的、前向き、後ろ向きのいずれの集計法で得られたデータを使っても、オッズ比は一定であるという性質があります。つまり、どんな調査法を使ってもオッズ比は推定できます。先に述べたように、リスク比は前向き調査でなければ推定できませんが、オッズ比はどんな調査でも推定できるわけです。ここが裁判の非常に重要な論点になりました。しかも、サリドマイド事件の場合、奇形児の発生率（P_A）はきわめて小さく、P_Bはもっと少なかったため、（$1－P_A$）も（$1－P_B$）もほとんど 1 に等しく、オッズ比に関する推計はほとんどリスク比に関する推計と同じと考えられます。サリドマイド事件は、統計の方法自体が論点になった希少な例ですが、後ろ向き調査の根拠が判例で認められたという点でも歴史に残る事例です。

もちろん後向き研究を行なう場合には、前向き研究や断面的研究と同等な母集団からの偏りのないサンプリングが達成できるかどうかに細心の注意を払う必要があります。後向き研究で研究対象とすべき母集団から正しくサンプリングが行われたか否かを巡っては、サリドマイド事件のように論争が起きることもあります。レンツ報告についても、サンプリングの妥当性が論点になったこともたしかです。

調査における無作為化についてもう一度整理しておきます。推測の対象となる集団（母集団）の一部をデータとして採取（標本の抽出）する場合には、どの対象をデータとするかは、ランダムに決めていきます。実際の調査において無作為抽出の方法論を確立させたのは、フィッシャーの次の世代のネーマンですが、それは 1930 年代の後半のことであり、比較的新しい手法です。理論が実際の社会調査で使われ定着していくためには、さらに時間が必要でしたが、そのいくつかの例を挙げておきます。

例　街頭調査の問題

アメリカでは 1940 年代から調査会社が存在したようですが、トルーマンとデューイの大統領選挙の時、ほとんどの世論調査会社は街頭調査を実施して、デユーイの優勢を伝えました。しかし、調査対象は街頭を歩いている有権者の集団という偏りがありました。そのため、ほとんどの調査会社は間違えた結論を出しましたが、一社だけ無作為抽出をしてトルーマンが優勢であると予測しました。このことがきっかけになって、無作為標本抽出の重要性が認識されるようになりました。意外に思われるかもしれませんが、サンプリングによって予測が逆転する事例はたくさんあります。

例　合格者の英語と数学との相関について

入試成績において、ほとんどの場合、英語と数学はかなり高い正の相関があります。一方、合格させることは偏ったサンプリングで、ランダムに決まるわけではありません。合格者は合計点でサンプリングされるために、負の相関に見えてしまいがちです。したがって、合格者の成績でみると、英語と数学には負の相関があり、英語と数学の成績はトレードオフの関係があると

いう誤解を与えてしまうこともあります。

例　SFC の一人っ子の比率と全世帯の一人っ子率

SFC の一人っ子の比率は全世帯の一人っ子率よりもかなり小さいと言われています。しかし、人をランダムに採取することと、世帯をランダムに採取することは異なります。ある大学のキャンパスで、1000 人の学生の中で一人っ子の比率をカウントしたとします。大学は学生をサンプリングしていますが、公的調査では世帯をサンプリングしています。つり、サンプリングの手法自体が違うわけです。

このように街頭調査の問題は現在でもあります。われわれが仮に、統計を使ってウソをつこうと思えば、たとえば寿司屋で寿司を食べている人に「あなたは寿司が好きですか」と聞くのが一番です。これは極端な例ですが、該当調査ではしばしばそういう問題があり、そういう調査を設計してしまえば、いかようにも変なデータを作ることができます。

ランダム化という操作は、基本的には、統計的推測の精度を上げるために行うのではなく、統計的方法の適用可能性を確保し、妥当性を向上させるために行う方法です。従って、ランダム化を導入しない調査や実験の方が、われわれに誤差の少ない情報を見せてくれる可能性はあります。たとえば、ある人物がキーパーソンであることがあらかじめ分かっている場合は、その人について調査する方法もあります。すなわち、有意抽出が必要な場合もあるわけです。その集団の特性や何を調べたいかによって、統計的手法を使うか、その他の調査手法を使うかを考えていくことが必要です。

４. 臨床試験はどうあるべきか？

4.1　実験として許されること

フィッシャーの実験計画法についてふれるために、ここでまた臨床試験に話を戻します。臨床試験はどうあるべきかについて、実験として許されることについて考えてみたいと思います。

まず、臨床試験が統計科学的であるべきだということは、今日コンセンサスになりました。一方、臨床試験の科学的説明に重点をおくか、臨床の現場における効果の評価という実践面を重視するか、つまり、仮説検証の効率か一般化可能性かのどちらを重視するかについては、医薬品の分野ではまだ論争中です。自動車や電気製品の開発など、許認可を伴わない産業界の分野では、市場での性能予測といった一般化可能性に重点が置かれていますが、医薬品開発は、学術と技術評価のはざま領域にあり、医学的な研究側面と同時に仮説検証の効率性の側面ももっています。つまり、説明的試験(Explanatory Trial)と実践的試験(Pragmatic Trial)の両面があるわけです。私自身は工学出身のため、当初は実践的試験を重視する立場でしたが、今では説明的試験の重要性も理解できるようになりました。

もう一つの問題は、プラセボコントロールはできるのか、ということです。説明的試験の最たるものである科学的仮説の検証のためには、ある薬がプラセボと比較してどのくらい有効性があるかを知る必要があります。

さらに、対象集団の選択と除外（不適格）の問題があります。医薬品の許認可のための実験をどういう対象で行なうか、逆に、どういう対象を不適格として除外するかに関しても、さまざまな論争があります。これについては、目標集団の範囲内でできるだけ広い患者を対象にすることがおおむね合意されてきています。ただ実際の臨床試験において、幼児や高齢者や妊婦を対象にしていいかどうかについては、いまだに議論されており、非常に慎重に配慮されています。

日本の過去の臨床試験の場合は、できるだけ広い患者を対象に、いろいろな病院や施設で実施するスタイルが多く採用されていました。それによってできるだけ一般化可能性を向上させることをねらっていました。それに対してアメリカの FDA は、まだ市販されていない薬を管理が不十分な施設で実験させることについてはきわめて慎重で、最良の施設で実施するという方針でした。ただし、最良の施設では一般化可能性は確保できない可能性があるので、複数の臨床試験をして再現性を確認するという立場をとっていました。

この点については日本とかなり論争になりました。日本の臨床試験は一つの病院では４～８例くらいとし、それを 100 施設くらいで実施していました。

論文を書くと、施設長がすべて著者になるので、1頁半くらい、すべて著者名が列記されているという、とんでもない論文も昔はありました。それも「ばらまき試験」など、アメリカに厳しく批判された点です。本当はどちらが望ましいかは非常に難しい問題で、最近はアメリカの論文でも、著者数80名というものがありましたので、アメリカでも判断はいろいろあるようです。

いずれにしても、臨床試験は、「たくさんの施設で1施設あたりの例数を少なく行うのが良いのか？」あるいは「少数の施設で、1施設あたりの例数を多く行うのが良いのか？」という問題に直面します。いずれも統計科学的な実験ですが、前者のほうがより統計的な実験に近く、後者のほうがより精密科学試験に近いと言えます。現在の規制では後者になっていますが、私は、国際的な論調に妥協することには反対の立場です。後者のほうが統計学的に根拠があると主張する研究者もいますが、これは統計的な根拠ではなく、むしろ科学的・経済的理由というべきです。

コックス(D. R. Cox) は、ピアソンの系譜を引く、現代統計学の第一人者ですが、彼がまだ若かった 1958 年に“Planning of Experiments”という本を著わしています。この本にはいろいろ興味深い指摘がなされています。たとえば、イギリスの農事試験場で、麦の品種改良をし、それを全国の農場に知らせたところ、スコットランドではその品種が全滅したそうです。それを受けて、1931年に“Student”は、結果の一般化可能性のない実験をしたからだとして、次のように述べています。

Thus the original experiments, carried out on well-farmed land, were definitely misleading when their conclusion were applied elsewhere.

ちなみに“Student”は、ギネスビールの生みの親であるゴセットのペンネームで、イギリスの品種管理の実践面での先駆者でした。ギネスビールとの知財関係があり、本名を明かすことができなかったようです。

私の立場をここで説明しておきたいと思います。私は工学部出身なので、一つの立場にこだわりすぎているのかもしれませんが、新薬の臨床試験について厚生労働省に対して発言する場合は、生産者と消費者の関係においては、消費者危険を保護する立場に立っています。統計的な検定を医薬品の許認可

に使うのは、聞いたことのない薬を市場に出す確率をある一定以下(５％以下)にしたいという意思があるからです。医薬品の許認可については、二相試験、三相試験で、少なくとも２回は５％検定の壁をくぐり抜けなければなりません。

また、実験計画法に対する立場としては、品質工学か古典的実験計画かという立場があります。私自身は、品質工学にこだわりはありますが、医薬品の分野では古典的実験計画の立場をとっています。さらに、多施設共同試験に対する立場では、施設内再現性か施設間再現性かという立場がありますが、私は後者を求めたいと考えています。その背景には、私が旧 JIS Z8402 分析・試験の許容差通則(1991 年)に関わっていたことも関係していると言えるかもしれません。

4.2 フィッシャーの「実験計画法」

フィッシャーは「実験計画法」において、無作為化の原則、繰り返しの原則、繰り返しの原則の３つの原則を示しています。

・無作為化の原則

すでに何度も繰り返して説明していますが、実験の中で無作為な割付けを行なうことです。それによって系統誤差を偶然誤差化できます。

・繰り返しの原則

当然、無作為化のサンプル数が小さいと偏りが生じてくるので、無作為化を実効性のあるものにするためには、適切な実験の繰り返しが必要になります。つまり、標本に一定の数を与えることによって、実験の誤差が評価できるようになるとともに、精度の向上も期待できます。

・局所管理（小分け）の原則

「局所管理」という言葉は一部では大変誤解を与えています。精密科学実験的な意味では、実験環境の管理と誤解されてしまいがちですが、品質管理の分野では「小分けの原則」と考えるべきです。これは精度の向上と一般化可能性を目的にしたものです。

フィッシャーの「実験計画法」の中では、「小分けの法則」は、ブロックの考え方で説明されています。正確には、繰り返しをブロックによって行なうべきだというものです。フィッシャーは農事試験場で実験を行なうにあたって、日あたりのいいブロック、水はけのいいブロックなど各ブロックに小分けして、興味のある対象をひとそろい全部入れました。そして、一つのブロックの中で一つの品種の反復実験をしても意義がないと主張しています。

ブロック数の意義

In increasing the number of blocks, however, we should have increased the area of the experiment, and it is probable that this increase in area, even if we had used the same number of larger plots, would itself have served to diminish the experimental error.

ブロック内反復の意義

If the area of the experiment were kept constant and the replication increased by using smaller plots we should only gain in precision

フィッシャーの原文は微妙なニュアンスがあり、なかなか難しい内容です。翻訳本も出されていて便利ですが、読むとかえって誤解するかもしれません。フィッシャーの本より、もう少し実験計画法について理解しやすいのが、『実験誤差の適切な選び方』(奥野・芳賀、1969 年)という本です。実は私は修士のときの 1981 年に、奥野先生から、この本を使った実験計画法の講義を受けました。その中で、次のような例題がありました。

〈例題〉

石油ストーブの芯の性能比較試験をする場合、どちらが良い実験計画か？

実験①－A、B それぞれの芯を取り付けた石油ストーブを 3 台ずつ用意し、各 1 回ずつ点火して CO 量を測る。すなわち、６台のストーブを用意して全部で６回実験する。(注：1 施設 1 例試験に該当する)

実験④－用いる石油ストーブはただ１台とし、これに A、B の２種類の芯を３つずつ用意して交互に取りつけ（ランダムな順で)、各１回ずつ点火して CO 量を測る。(注：単一施設試験に該当する)

奥野先生は、このどちらの方法がよいかを質問され、たまたま私が指名されました。私は④と回答しました。その理由を尋ねられたので、私は「評価したいのは芯ですよね」と質問し、「そうだよ」という答えをいただきました。そこで私は、芯の差を評価するのであれば、ストーブをたくさんにすればストーブのバラツキが出て、実験の誤差が大きくなるが、④なら、ストーブのバラツキは出てこないので、芯の差がきれいに出てくると思います、と答えました。

すると奥野先生は、「それが素人のあさはかさ」と言われました。それは今でも大変鮮明に覚えています。この問題に対する奥野先生の答えは次のように書かれています。

ここで、『誤差分散』の大きさだけからいえば、④の方が①より小さく、すぐれていることになる。
しかし、結論の普遍性という立場からみると、④の結論は用いたその 1 台のストーブに限定されるのに対し、①では『ストーブ間の個体差』を『誤差』に入れているので、広い範囲にその結論を適用しても大過ないであろう。この観点からは（そして実際家の立場からは）①の方がすぐれているというべきである。

私が回答したように、誤差のバラツキだけで言えば、④のほうが①より小さくなります。その意味では私の答えはまちがいではありません。ところが、結果の普遍性、一般化可能性の立場からは、④は１台のストーブに限定されてしまいます。それに対して①は、ストーブの個体差を考慮しても芯の差が出てくるのであれば、広い範囲に結論を適用できるという点で、①のほうがすぐれています。つまり、いろいろな環境条件の中で実験を実施したほうが、一般化可能性のある結論を導き出すことができるのです。産業実験はそうあるべきだと言えます。私が主張した④は、実験室内での再現性を考えればいい精密科学的実験の立場だったわけです。

このように①のほうが実験としてはいいのですが、誤差は大きくなります。それを解決したのが、フィッシャー自身の回答です。フィッシャーは、実験

はブロックの中で繰り返すのがよいと主張しました。そして実験⑤として、３台の石油ストーブを用意し、それぞれに２種類の芯A、Bを交互に取りつけて(ランダムな順で)、各１回ずつ点火しCO量を測りました。これは、ランダマイズド・ブロック(乱塊法)と呼ばれている実験で、１施設１群１例試験にあたります。

統計的解析によって「ストーブ間の差」は「誤差」要因から除去されます。無作為に採用した３台のストーブについての結論ですから、それだけ適用範囲が広いことになります。これがフィッシャーの考えた、「小分けの原則」です。統計的実験の歴史は、乱塊法の積極利用の歴史でもあります。科学的精密実験とは発想が異なることに注意する必要があります。

4.3　精密科学実験 vs 統計的実験再び

コックスは1958年に、次のような指摘をしています。

These considerations are rather less important in purely scientific work, where the best thing is usually try to gain through insight into some very special situation rather than to obtain in one experiment a wide range of conclusions

このように、一般可能性に関する議論(These considerations)は、純粋な科学的研究(purely scientific work)においては重視されていないと主張しています。私は統計的実験を産業界で使う立場でも仕事をしてきましたが、やはり産業界と学術研究界とではここに差があると思います。アカデミアにおいては室内再現性が重視され、一般可能性は次の段階の課題ととらえられています。

このように、新薬許認可のための臨床試験は、“purely scientific work”かどうかという点にゆらぎがあります。それは純粋な科学的研究であると主張する研究者もいます。しかし、その成果をもとに新薬は市場に出るのですから、私はそのことを重視します。FDAは科学的な実験を２回行なうとしていますが、私は、方針の間違った実験を繰り返しても普遍性確保にとっては非

効率だと主張しています。

交互作用解析のために 1 施設でたくさんの実験をする必要があるという主張に対して、われわれは、それは「実験計画法」の趣旨からしておかしいと反論しました。それに対してアメリカは、施設の中にはおかしな施設もあるから、日本のように症例数が少ない実験では、その施設がおかしいかどうか判断できないと指摘しました。おかしな施設を排除するために、たくさんの症例数を確保するという論点はどうなのか。臨床試験は薬の特性を調べるものです。交互作用というのは、薬と施設の間の特徴づけですが、そういう現象があるかどうかを調べるだけだったら、むしろ施設数は増やしたほうが検出力は増大します。

この点についても、私はアメリカに対して相当反論してきました。ただし、日本国内の専門家でもアメリカの主張を指示する立場もあり、私一人苦戦してきたという状況でした。しかし統計的な原理や手法としては、私の主張はまちがっていません。統計の国際標準化規格 ISO5725 は、まさにこの問題についての規格ですから、そもそも医薬品分野が ISO5725 に反する臨床試験をするのは疑問です。

コックスも 1997 年になって、共同施設の共同実験に関してさまざまな手法が発達し、ISO 5725 のような国際規格化が成立したと述べています。

In many physical science contexts, especially those connected with industrial applications, there is considerable emphasis on international standardization of measurement method. There is in some fields emphasis on the design and analysis of interpersonal and interlaboratory studies of measuring techniques. These are regrettably less common in a medical context.

おそらく多くの人は、統計的な手法について、国際標準があることすら知らないでしょう。医薬品業界に関係する統計の中で、私のような主張をする統計家はほとんどいません。しかし、このコックスの主張に見られるように、国際的な統計学の中で孤立しているわけではありません。

4.4　ISO5725 の主要な推奨事項記述

これについては、簡単に紹介するにとどめます。施設数は８～15 はないとまともな実験はできないことを根拠をあげて指摘しています。逆に 20 施設以上多くしても実験精度は上がらないし、５施設以下だと正確なデータは得られないとしています。さらに、1 施設当たり 1 水準当たりで n を 2 以上にしても得るものはほとんどないとも指摘しています。

私は、1 施設でたくさんの実験をしたほうがコストが安くなるという経済的理由なら許容できますが、統計的な根拠としてそのことが正当化されるのは絶対許せないという立場をとってきました。私と同じ主張をされているのは、東京大学工学部の広津千尋先生です。

これについてはさまざまな議論がなされていますが、最低、バラエティに富む８施設が参加すれば実験はなんとかうまくいくだろうという見解が落とし所だと考えています。日本はもともと 100 施設くらいで実験していたので、国際的ガイドラインが発行され、10 施設くらいで実験するようになったため、事態は改善されたということになっています。

ICH E9 は日米欧の３極で合意して発行されましたが、正確には、日米欧の６極と呼んでいます。なぜなら、３極の中に、それぞれ許認可側(日本の厚生労働省、アメリカの FDA)、申請側(製薬会社)が入っているからです。日本の統計解析の旧ガイドラインは、世界で初めての臨床試験分野のガイドラインでしたが、その後、アメリカ、ヨーロッパでガイドラインを策定することになり、新ガイドラインができたわけです。結果として、日本の考え方は通りませんでした。

日本でガイドラインを作ったときは、医師と厚労省が参加し、製薬会社はいっさい入りませんでした。ガイドラインの作成にあたっては、生産者側である製薬会社が参加するのは正しい姿かとは思いますが、われわれが当時日本のガイドラインを作成しているときは、消費者側が枠組みを作り、精神者はその中でベストを尽くせばいいと考えていました。これは、ロシアなどではプライオリティ・プリンシプルと呼ばれています。パイを正確に切るのは行政側で、その中でベストなものを選ぶのが生産者側という発想です。ICH E9

は、ルールを生産者と消費者双方で決めるという発想です。どちらがいいかは、統計の中ではなかなか決着のつかない問題です。

ICH E９文書が発行され、乱塊法実験は制度上行えなくなりました。われわれとしては、次善の策として、１施設の中でブロックを形成することを考えました。施設内の評価医師を新たにブロック因子として、例えば１医師当たり１群２例、計４例を割り付ける試験で、１施設を５ブロック以上で構成します。

実際、臨床試験では 1 例あたり非常にお金がかかります。したがって症例数はできるだけ節約して効率的に実験を行ない、よい結果を出したいと考えるでしょう。総症例数が少なければ管理のコストも低くおさえることができます。ですから、経済的な観点という理由であれば、それは合理的な判断と言えます。結果として、現時点ではプラグマティック・トライアルが減りました。日本はもともと第３相試験では、プラグマティック・トライアルを採用していたので、それをしなくなったのは、個人的には非常に不満でした。患者や医師にできるだけ様々なタイプが参入でき、日常診療を再現するような工夫が、市販を前にした最終関門としての第３相試験の試験計画では必要です。それが日本のコンセンサスであったはずなのに、あっという間に消失してしまったのは驚きです。

医薬統計分野の第一人者で、オックスフォード大学教授のピーター・アーミテージは、1996 年に、臨床試験のあり方全般について、次のように述べています。彼は、生物統計学の学会長をつとめるとともに、クリニカル・サイエンスの学会長だったこともあり、その指摘は傾聴に値すると思います。

Most trials exhibit both explanatory and pragmatic features, but in practice the pragmatic attitude tends to dominate

なお、京都大学医学部教授の佐藤俊哉さんの『宇宙怪人しまりす医療統計を学ぶ』(岩波書店)は、医療統計に関心のない方でも、一読すると、実験計画法をはじめ、いろいろなことがよく分かる大変興味深い本ですので、関心のある方には一読をおすすめします。

【参考資料】

椿・藤田・佐藤編：これからの臨床試験，朝倉書店

中谷　科学の方法，岩波書店

GCP，厚生省

ICH　E9，厚生省

旧臨床試験のための統計解析ガイドライン，厚生省

椿，藤田，佐藤，誰がための臨床試験，統計数理

Armitage and Berry，椿訳，医学研究のための統計的方法，サイエンティスト社

Pocock,コントローラ委員会訳，臨床試験，篠原書店

歴史的資料

佐藤(1977)薬剤の評価(有用性)に対する考え方の変遷,内科,39 巻 6 号,913-916

「佐久間昭の世界」，サイエンティスト社

椿，藤田，佐藤（1998）誰がための臨床統計，統計数理，46 巻 1 号，97-115

５.倫理について

5.1 「技術倫理」の重要性

私は倫理について専門ではありませんが、1988 年か 89 年に、アメリカのミズーリ―大学に交流のために行き、日本の品質管理の教育、現状について話していたとき、先方の責任者から、日本では技術倫理について教育しているのかと聞かれました。当時は何もしていないと答えたところ、技術倫理についてはぜひ教育すべきだと強く主張されました。そして『ハリスの倫理学』(翻訳：日本技術史協会)という本を推薦されました。その本では大変重要なことが書かれていて、技術倫理について学校教育で学ばなければならないとつくづく感じました。

この中では、「技術者は、専門職として有能な領域を越えて、その専門職の重みによって公衆への発言をしてはならない」と指摘されています。私は統計家ですから、倫理の話をするのはためらう面もありますが、先ほど紹介した佐藤俊哉さんも、科学的であることと倫理的であることには、きわめて重要な接点があるとおっしゃっているので、あえてとりあげてみたいと思いま

す。

倫理は英語では、ethics あるいは moral だと思います。しかしベートーベンは耳が聞こえなくなり、作曲家として苦悩した時期に「ハイリゲンシュタートの遺書」をしたためていますが、そこでは、自分が危機から救われたのは、virtue によってであると書かれています。もちろん原文はドイツ語ですから、英訳すると virtue になります。「ハイリゲンシュタートの遺書」は日本語にも翻訳されていて、virtue は道徳と訳されています。

カール・ピアソンが「科学の文法」を作った際、その前提として、『自由思想の倫理(The Ethic of Freethought)』を書きましたが、その中で、闘いの倫理として奮い立たせる原動力としての科学に言及しています。そして、プロフェッショナルの倫理として virtue をあげています。moral や ethic は、皆で守らなければならない規範の意味ですが、virtue はベストを尽くして何かをするというニュアンスです。

そういう意味で、ベートーベンが音楽の世界でベストを尽くす原動力としての virtue に救われたということは理解できます。それに対して技術倫理における virtue は、「ベストを尽くして当該分野において人知が及ぶ領域を拡大させる」という意味になると思います。

5.2 『ハリスの倫理学』に見る倫理規定

『ハリスの倫理学』(1995 年版)では、科学技術者の努力目標として、次のような規準をあげています。

・比較規準
　通常われわれが受け入れているリスクより大きくないこと
・正常性規準
　自然状態と同じ程度であること。環境基準などはこれが求められます。
・最適汚染削減の規準
　コスト・ベネフィット分析上の効用関数が最適となること
・最大保護の基準
　人体影響の可能性のあるリスクが全て排除されること

・証明可能危害の規準

人体影響が立証されているリスクが全て排除されること

上記の規準は、それぞれの立場が違うとまったくかみあわないことになります。実際の例として、数年前に、インフルエンザの治療薬タミフルの副作用の問題がありました。タミフルを飲んだ副作用で飛び降り自殺をした例があるとされています。データで見る限り、可能性のあるリスクではありました。一方で、それが実証リスクかどうかについては、数年前には実証されてはいませんでした。厚労省はその段階で、高校生・中学生以下にはタミフルの投与を禁止しましたが、アカデミアの世界では論争が続きました。最適汚染削減の規準では、社会的な期待効用が最適化される行動を求めることが集団倫理の上位概念となっています。こういう議論は、1990 年代から少しずつ始まりましたが、近年は、ハーバード大学の「白熱教室」でもさかんに取り上げているため、われわれにとってもなじみ深いものになってきました。

これまでのことから明らかなように、倫理については、次の２つの立場からの主義・主張があります。

①社会の倫理(功利主義)

最適汚染削減の規準を導くもので、個人の行為はその影響をうける者に最大の効用を与えるのが正しいという立場です。

②個人の倫理(Respect for Person)

モラル体現者(自分自身で自律的に意思決定できる存在=人間)としての各個人を平等に尊重するのが正しい行為という立場です。

実際は、この２つの立場のバランスをどうとるかにつきると思います。『ハリスの倫理学』(1995 年版)では、危害度規準(degree of harm criterion)を設け、「人体影響の差し迫った脅威が、経済的に合理的な費用で危険性のないものになっていること」と定めています。特に、環境問題では、この規準が適応されるべきでしょう。これが 1995 年当時のハリスの考え方でした。「1995 年当時」と限定したのには理由があります。というのは、倫理はどの時代でも不変であるわけではなく、時代とともに進化していくからです。後に、医

療関係の倫理規定である「ヘルシンキ宣言」についてふれますが、数年に１回改訂されて、少しずつ進化しています。

その進化の一例として、ハリスが紹介している土木技術者の倫理を取り上げてみましょう。アメリカ土木技術者協会(ASCE)規定(1977 年)では、「技術者は、生活の質を高めるように環境を改善すべく活動することが望ましい(should)」となっています。1980 年には「土木技術者は、自分たちの努力が環境に及ぼす影響を認識すること、および、選択された設計が環境に及ぼす影響を依頼者に報告すること」という方針声明が出されます。1983 年には、「技術者がサービスを行う際には、世界の資源、ならびに、自然のままおよび人手が加えられた環境を現在及び将来の世代の受益のために、節約して使用しなければならない(shall)」と提案しました。

ここで少し補足しておきますと、倫理規定などでは一般的に、should は「勧告」などに、shall は「要請」などに使われます。経済産業省は、ISO の翻訳ルールを決めて、should は「望ましい」、shall は「しなければならない」と訳すよう統一しています。どちらの言葉が用いられているかについては、私も大変神経を使いながら読むようにしています。全般的に、多くの倫理規定や標準規定などでは、最初は should が多く用いられ、しだいに shall や must に変わっていくというのが通例です。人間の倫理的な行動は、その社会の経済的発展状況によって大きな差がありますから、最初は「望ましい」から始まり、しだいに「しなければならない」という規定に変わってくるわけです。

実際に、1983 年に倫理規定を作る側が出した提案は、1984 年に理事会全員が一致して否決しました。should を shall に変更することを拒否したわけです。1996 年の改訂規定で、やっと shall が承認され、「技術者は、公衆の安全・健康・福利を最優先しなければならない」「技術者の職務行動は継続的発展の原理に従っていなければならない」となりました。

5.3　ヘルシンキ宣言の概要

以上のことを念頭に置きながら、ヘルシンキ宣言を紹介していきます。すでに述べたように、ヘルシンキ宣言はヒトを対象とする医学研究の倫理的原

則です。この原則をきちんと満たしているかどうかをチェックするために、多くの医療関係機関では倫理委員会を発足させています。統計数理研究所にもあり、私が委員長をつとめています。ヘルシンキ宣言は日本医師会ホームページには日本語で掲載されていますが、原文は世界医師会にあります。
http://www.wma.net/e/policy/b3.htm

そもそもは、1964年6月にフィンランドのヘルシンキで開催された、第18回 WMA 総会で採択されたため、この名前がつけられました。倫理規定にもかかわらず永遠不変のものではなく、今日まで繰り返し修正が続いています。特に1996年以後は、臨床試験に特化したさまざまな修正が行なわれ、プラセボを使うことに関する倫理規定が増加しました。もともとの「ヘルシンキ宣言」を忠実に読むと、標準治療が確立している治療についてはプラセボは使えないという既定だったのですが、プラセボを使えるための既定に修正されています。もっとも最近は2008年に改正されています。

「ヘルシンキ宣言」をすべて詳しく紹介はしませんが、序言のいくつかをとりあげてみましょう。

序言1

The World Medical Association has developed the Declaration of Helsinki as a statement of ethical principles for medical research involving human subjects, including research on identifiable human material and data.
The Declaration is intended to be read as a whole and each of its constituent paragraphs **should** not be applied without consideration of all other relevant paragraphs.

世界医師会（WMA）は、個人を特定できるヒト由来の試料およびデータの研究を含む、人間を対象とする医学研究の倫理的原則として、ヘルシンキ宣言を発展させてきた。
本宣言は、総合的に解釈されることを意図したものであり、各項目は他のすべての関連項目を考慮に入れず適応されるべきではない（注：本来は**望ましくない**）。

日本医師会の翻訳は、should を「べき」としています。この指針をもとに、厚生労働省や文部科学省が指針を作っているために、日本の指針は事実上、

must になっています。各国ではまだ実現できないため、should で書かれている部分まで日本では「べき」と解釈され、負担が大きくなっていると言えます。その意味では、翻訳もきわめて重要です。

序言４(旧３)：Shall は要請

The Declaration of Geneva of the World Medical Association binds the physician with the words, "The health of my patient will be my first consideration," and the International Code of Medical Ethics declares that, "A physician **shall** act only in the patient's interest when providing medical care which might have the effect of weakening the physical and mental condition of the patient."

世界医師会のジュネーブ宣言は、「私の患者の健康を私の第一の関心事とする」ことを医師に義務づけ、また医の倫理の国際綱領は、「医師は患者の身体的及び精神的な状態を弱める影響をもつ可能性のある医療に際しては、患者の利益のためにのみ行動すべきである（注：本来は**しなければならない**)」と宣言している。

ここでは shall が使われていますから、医師会の翻訳よりは、もっとずっと強いニュアンスになります。

倫理規定がどう変わっていくかという点で、「ヘルシンキ宣言」は大変重要なので、2008 年の改訂で、どんな内容が追加されているかにも大きな関心があります。序言５では臨床試験の必要性について述べており、医学研究で過小評価されている研究対象にも、研究参加への適切なアクセスの機会が提供されることが望ましいと書かれています。なお、序言６の被験者保護では、2008 年の改訂で初めて、should が shall になりました。

序言６：被験者保護が推奨事項から要請に

In medical research on human subjects, considerations related to the well-being of the human subject must(**今般改訂により should から変更**) take precedence over all other interests

人間を対象とする医学研究においては、個々の研究被験者の福祉（注：良い状態のことではないか？）が他のすべての利益よりも優先されなければならない。

また序言７では、治療方法の継続的改善について書かれています。これは

リスクの問題に関わってきます。リスクは語源的には、一歩前に進むというニュアンスもあり、人間はリスクをとらなければ進歩しないという面もあります。医薬品の開発や治療行為の改善は積極的に行なうべきですが、同時にリスクも伴います。現在の治療水準に満足していれば、新薬の開発や治療行為の改善は必要ないわけです。佐久間昭先生もよくおっしゃっているように、「くすりはリスク」です。序言８においても「医学の実践および医学研究においては、ほとんどの治療行為にリスクと負担が伴う」と明記されています。

そういう中で、序言９では「倫理基準の遵守」について、「医学研究は、すべての人間に対する尊敬を深め、その健康と権利を擁護するための倫理基準に従わなければならない」と述べ、弱者に対する権利の保護や同意のとり方についても言及しています。さらに、序言 10 では、他の基準の遵守と基準への要請について、次のように述べています。

序言 10：他の基準の遵守と基準への要請

Physicians **should** consider the ethical, legal and regulatory norms and standards for research involving human subjects in their own countries as well as applicable international norms and standards.
No national or international ethical, legal or regulatory requirement **should** reduce or eliminate any of the protections for research subjects set forth in this Declaration.

医師は、適用される国際的規範および基準はもとより、人間を対象とする研究に関する自国の倫理、法律および規制上の規範ならびに基準を考慮するべきである（注：本来は、**ことが望ましい**）。
いかなる自国あるいは国際的な倫理、法律、または規制上の要請も、この宣言が示す研究被験者に対する保護を弱めたり、撤廃するべきではない（注：本来は、**ことは望ましくない**）。

国によっては、「ヘルシンキ宣言」が守られていない部分もありますので、should で書かれていますが、医師会が「べきである」と訳しているので、日本政府は、忠実にこれを守っています。

その他、「ヘルシンキ宣言」では、多くの原則が定められていますので、2008

年に追加された新たな原則を中心に、主要なものを紹介します。

すべての医学研究のための諸原則

◎基本原則 11：個人情報の保護

ここで、個人情報保護条項が新たに追加されました。「研究被験者の生命、健康、尊厳、完全無欠性、自己決定権、プライバシーおよび個人情報の秘密を守ることは、医学研究に参加する医師の責務である」と定めています。

◎基本原則 12：やみくもな実験の禁止

ここでは「人間を対象とする医学研究は、科学的文献の十分な知識、関連性のある他の情報源および十分な実験、ならびに適切な場合には動物実験に基づき、一般的に受け入れられた科学的原則に従わなければならない」「研究に使用される動物の福祉は尊重されなければならない」と定めています。

この基本原則があるため、われわれ統計家のように医学関係者以外も倫理委員会に参加することになっています。倫理委員会では、実験計画に対して科学的・倫理的に妥当であるかどうかを評価します。

◎基本原則 13：環境保護

ここでは、「環境に悪影響を及ぼすおそれのある医学研究を実施する際には、適切な注意が必要である」と訳されていますが、原文では、must be となっていますので、「注意をしなければならない」と考えるべきです。

◎基本原則 14：プロトコルへの明示が勧告から要請へ

この原則は、今回の 2008 年改訂で、should から shall に変わりました。

The design and performance of each research study involving human subjects **must** (これまで should)be clearly described in a research protocol.

人間を対象とする各研究の計画と作業内容は、研究計画書の中に明示されていなければならない。

これによって、人間を対象とする各研究の計画と作業内容を研究計画書の中に明示しなければならなくなりました。ただし、計画書の中に何をもりこ

むかについては、まだ should 規定が多いのですが、日本では「べき」と訳されているため、以下のすべてを遵守していることになります。

・倫理的配慮

研究計画書は、関連する倫理的配慮に関する言明を含み、また本宣言の原則にどのように対応しているかを示すべきである（ことが望ましい)。

・被験者の補償に関する条項

計画書は、資金提供、スポンサー、研究組織との関わり、その他起こり得る利益相反、被験者に対する報奨ならびに研究に参加した結果として損害を受けた被験者の治療および／または補償の条項に関する情報を含むべきである（ことが望ましい)。

・被験者の研究後のアクセス

この計画書には、その研究の中で有益であると同定された治療行為に対する研究被験者の研究後のアクセス、または他の適切な治療あるいは利益に対するアクセスに関する取り決めが記載されるべきである（ことが望ましい)。

◎基本原則 15：独立倫理委員会の審査とモニタリング

これは以前から確立されている原則ですが、研究計画書は、検討、意見、指導および承認を得るため、研究開始前に研究倫理委員会に提出しなければなりません。審査委員会は、当然利害相反のない人たちで構成されていなければなりません。委員会は、適用される国際的規範や基準はもとより、研究が実施される国々の法律と規制を考慮しなければなりませんが、それらによって「ヘルシンキ宣言」が定めている研究被験者に対する保護を弱めたり、撤廃することは許されません。

同時に、倫理委員会は、進行中の研究を監視(モニタリング)する権利を持っています。日本では、国が直轄している大規模な臨床試験では、倫理モニタリング委員会を設置しています。研究者は委員会に対して、監視情報、とくに重篤な有害事象に関する情報を提供しなければならないと定められています。たとえば、一つのグループで非常に特異的な症状が出ているとき、実験を中止するなどの判断をすることもあります。また、委員会の審議と承認

を得ずに計画書を変更することはできません。たとえば臨床試験の症例数が集まらない場合には、臨床試験時期の改定などを倫理委員会にはかる必要があります。

◎基本原則 16：科学的力量のある専門家に支えられた研究

ここでは「人間を対象とする医学研究を行うのは、適正な科学的訓練と資格を有する個人でなければならない」と定められています。すでに述べたように、欧米では、この中に統計家が入り、科学的に重要な役割を果たしているのが通常です。当然、患者あるいは健康なボランティアに関する研究は、能力があり適切な資格を有する医師もしくは他の医療専門職による監督が必要であることは言うまでもありません。被験者の保護責任は常に医師あるいは他の医療専門職にあり、被験者が同意を与えた場合でも、決してその被験者に責任を帰することがあってはなりません。

◎基本原則 17：弱者対象研究(新独立条項)

これは新たに設定された条項で、「不利な立場または脆弱な人々あるいは地域社会を対象とする医学研究は、研究がその集団または地域の健康上の必要性と優先事項に応えるものであり、かつその集団または地域が研究結果から利益を得る可能性がある場合に限り正当化される」と定めています。

これは、わが国にとって今後、非常に大きなテーマとなります。福島第一原発事故後、福島地域で公衆衛生学的な研究が行われる可能性は十分にありますが、研究者個人の関心のために研究できるかという問題があります。研究が福島地域の健康上の必要と優先事項に応える場合、あるいは地域がその研究結果から利益が得られる可能性がある場合に限り正当化されます。逆に言えば、学術上の興味だけで調査研究することは許されません。

◎基本原則 18：事前のリスク評価(勧告から要請へ)

これによって、事前のリスク評価が義務づけられました。臨床試験を行なう場合、予想されるメリットとリスクについて、リスク評価の観点から事前に調べなければなりません。

◎基本原則 19：臨床試験の事前登録(新規要請)

ここでは、「すべての臨床試験は、最初の被験者を募集する前に、一般的にアクセス可能なデータベースに登録されなければならない」と定められています。これはまだ、厚生労働省の臨床試験のガイドラインに反映されていない原則ですが、登録機関は東大医学部に設置されています。

◎基本原則 20（旧 17）：研究の中止・不参加（勧告から要請）

今回の改訂で、「医師は、内在するリスクが十分に評価され、かつそのリスクを適切に管理できることを確信できない限り、人間を対象とする研究に関与することはできない」となりました。また、「医師は潜在的な利益よりもリスクが高いと判断される場合、または有効かつ利益のある結果の決定的証拠が得られた場合は、直ちに研究を中止しなければならない」と、いずれも勧告より強いニュアンスの要請になりました。

一般的に、二重盲検比較試験をしている場合、自分がどちらの群に割り付けられているか分かりません。この原則を守るためには、第三者委員会を設置し、ブラインドの実態をモニタリングしている必要があります。途中で中止するためには、何回か結果をチェックしなければいけませんが、どの程度のレベルなら中止していいかは統計的な手法で判断せざるをえません。この判断基準は統計家が作らなければなりません。

◎基本原則 22（旧 20）：ボランティアと同意（must 要件）

これは以前から「ヘルシンキ宣言」の重要な要件です。まず、医学研究への被験者としての参加は、判断能力のある個人による自発的なものでなければならないと定めています。第１相試験で健常な成人男性という規定がある場合など、アルバイト感覚で参加するケースがあると言われています。いずれにしても、本人の自由な承諾なしに研究へ登録してはなりません。

◎基本原則 23（旧 21）被験者のプライバシー保護（勧告から要請に）

これも新たに重視された原則で、「研究被験者のプライバシーおよび個人情報の秘密を守るため、ならびに被験者の肉体的、精神的および社会的完全無欠性に対する研究の影響を最小限にとどめるために、あらゆる予防策を講じなければならない」としています。

◎基本原則 24：十分な説明の上での同意と撤回

これは従来から、臨床試験では重要な原則で、いわゆるインフォームド・コンセントについてです。主な内容は次のとおりです。

・判断能力のある人間を対象とする医学研究において、それぞれの被験者候補は、目的、方法、資金源、起こりうる利益相反、研究者の関連組織との関わり、研究によって期待される利益と起こりうるリスク、ならびに研究に伴いうる不快な状態、その他研究に関するすべての側面について、十分に説明されなければならない。
・被験者候補は、いつでも不利益を受けることなしに、研究参加を拒否するか、または参加の同意を撤回する権利のあることを知らされなければならない。
・被験者候補がその情報を理解したことを確認したうえで、医師または他の適切な有資格者は、被験者候補の自由意思によるインフォームド・コンセントを、望ましくは文書で求めなければならない。

◎基本原則 25：個人識別可能なデータ

個人識別可能なデータの扱いについても、次のような原則が定められています。ゲノム研究ではこの原則に関わる問題が出てくると思われます。

・個人を特定しうるヒト由来の試料またはデータを使用する医学研究に関しては、医師は収集、分析、保存および／または再利用に対する同意を通常求めなければならない。
・このような研究には、同意を得ることが不可能であるか非現実的である場合、または研究の有効性に脅威を与える場合があり得る。このような状況下の研究は、研究倫理委員会の審議と承認を得た後にのみ行うことができる。

◎基本原則 26（旧 23、24）：インフォームド・コンセントへの勧告

医師に頼まれれば断りにくいという関係があるため、インフォームド・コンセントについて「研究参加へのインフォームド・コンセントを求める場合、医師は、被験者候補が医師に依存した関係にあるか否か、または強制の下に同意するおそれがあるか否かについて、特別に注意ことが望ましい」としています。「また、このような状況下では、インフォームド・コンセントは、そ

のような関係とは完全に独立した、適切な有資格者によって求められるひとが望ましい」としています。さらに、本人が意識不明状態のときのみ、代理人の承認もありうると定めています。

◎基本原則 30（旧 27）：研究結果の公表（勧告）

ここでは、「著者、編集者および発行者はすべて、研究結果の公刊に倫理的責務を負っている」と定めています。著者には、人間を対象とする研究の結果を一般的に公表する義務があり、報告書の完全性と正確性に説明責任を負っています。また、倫理的報告に関する容認されたガイドラインを遵守することが望ましいとしています。

また勧告ではありますが、「消極的結果および結論に達しない結果も積極的結果と同様に、公刊または他の方法で一般に公表されることが望ましい」と定めています。つまり、当初の仮説に基づいた実験の結果がうまく出なかった場合でも、その結果の公表を求めているわけですが、大変難しいのも事実です。

学術論文であれば、ネガティブな結果はまずアクセプトされないでしょう。研究者にとっては失敗した実験結果の公表には大変抵抗がありますが、臨床試験の場合、それでも公表が求められます。これは、いわゆるパブリケーション・バイアスを避けるためです。医学に限らず、社会ではポジティブな結果が出たもののみがデータとして流通し、ネガティブなデータは知られないままになります。薬の場合、世界各地でいろいろな臨床試験が行なわれている可能性があります。その中で、ポジティブな結果だけが公表されるとすれば、その薬の効果は過大に評価されることになります。それを避けるための原則です。そもそも医学分野では、「ヘルシンキ宣言」に違反した実験は公刊の権利はないと見なされています。

◎追加原則 32（旧 29）：通称プラセボ条項

「ヘルシンキ宣言」のもともとの原則は、「新しい治療行為の利益、リスク、負担および有効性は、現在最善と証明されている治療行為と比較考慮されなければならない」というものでした。つまり、被験者は臨床試験に参加するにしても、最善の治療を受けることが大原則だったのです。ところが、きわ

めて珍しい例ですが、開発途上国ではなく、FDA の立場を許容するため、「ただし、以下の場合にはプラセボの使用または無治療が認められる」という条件として、次の２つが盛り込まれました。

①現在証明された治療行為が存在しない研究の場合
②やむを得ない科学的に健全な方法論的理由により、プラセボ使用が、その治療行為の有効性あるいは安全性を決定するために必要であり、かつプラセボ治療または無治療となる患者に重篤または回復できない損害のリスクが生じないと考えられる場合

①については誰でも納得できるでしょう。②については、個人の倫理と集団の倫理の観点から、新薬の臨床試験の場合は、集団にリスクを与えるよりは、個人に復元可能なリスクを許容するほうがまだましだという判断なのでしょう。日本にとっては、こういう条項は本来必要ではなく、最善の治療法による臨床試験を行なっていました。また日本は国民皆保険のため標準治療は受けられますが、アメリカはそうではありません。臨床試験に参加すれば、確率 50%でプラセボ群になるかもしれないが、最先端の治療を受けられる可能性もあるという期待で参加する人もいます。このように文化や制度の差を無視して、プラセボ条項を盛り込むことについては、当初の「ヘルシンキ宣言」の理念からして、私はきわめて違和感を感じています。

「ヘルシンキ宣言」自身もこの点については非常にゆらいでいて、「この手法の乱用を避けるために十分な配慮が必要である」という文章も入っています。他に比べると、この条項は政治的な要因から生まれたもので、きわめて混乱の中にあると言えます。

生命科学などを研究されている方は、最新の「ヘルシンキ宣言」を確認するようにしてください。日本では「厚生労働省臨床研究に関する倫理指針」平成 20 年改訂が最新だと思います。「ヘルシンキ宣言」に準拠していますが、先に指摘したように、should は医師会訳に従い全て shall と解釈されているという問題もあります。

倫理委員会では、一般的には、「厚生労働省臨床研究に関する倫理指針」か

文部科学省・厚生労働省の「疫学研究に関する倫理指針」を使って判断します。人に対する介入行為があるかどうかが判断基準になっており、介入行為がある場合は前者、ない場合は後者が使われます。ただし例外があり、栄養食品など食品に関する無作為介入研究の場合は、厚労省の指針でなくてもよいとされています。

また、文部科学省・厚生労働省の「疫学研究に関する倫理指針」には、インフォームド・コンセントに関しても、簡略化もしくは免除規定があります。個人単位の介入では免除はありませんが、集団単位の介入では，簡略化が許容される場合もあります。たとえば、糖尿病予防の調査で、地域ごとに、運動、健康飲料の服用など、方法が異なる場合などです。こういう場合には、ポスターなどで参加を呼びかければ、インフォームド・コンセントがあったと見なされます。

5.4　医薬品の安全性確保のために

最後に医薬品の安全性の確保についてふれておきたいと思います。
安全性について統計学が関わる場合のことを考えてみます。医薬品については、承認時には、この薬には○○に効くという保証の論理を採用します。しかし、市販後の安全性については、エビデンスはまだ明確には証明されていなくても、それなりに警告を始める必要があります。ここは重要な点です。

承認の際は、この薬はここが危ない、そこが危ないかもしれないと多重性の論理はあまり使いません。しかし安全性の場合は、ここが危ないかもしれないというものが出てくるほうが望ましいわけです。先ほどのタミフルの例と同様です。一方で、情報収集の仕組みを作り、エビデンスのレベルが高まっていけば、ある段階でリスクの囲い込みをしていきます。その後、さらに統計的な検定で保証されるレベルになれば、禁止などの是正措置を講じてきます。

つまり、有効性と安全性の判断基準は異なっていて、決して対照ではありません。有効性の場合は、ここが効いているというだけの“つまみ食い”は許されませんが、安全性の場合は“つまみ食い”が許されます。なぜなら、

有効性はオア・ロジックで、A　or B or……のどれかが有効である場合、全体でどうかという観点で考えていきます。どこかが有効であれば、製薬会社の申請を認めてもいいかもしれません。しかし、安全性はアンド・ロジックで、どれ一つ崩れても困るわけです。どの部分の有害事象についても保証しなければなりません。このように、両者はロジックが異なるので、それによって用いられる統計の方法も異なってきます。社会とつきあう場合は、その分野の論理を汲み取り、適切な方法論を設計していかなければなりません。それが統計家のミッションと考えています。

マーケティングの統計科学 *

照井伸彦 †

2025 年 8 月 25 日

概要

本稿は、マーケティング分野における統計科学について学ぶことを目的とし、下記の構成でマーケティングにおける統計科学の役割とその応用について概観する。具体的には下記の事項について解説する。(i) マーケティングと統計科学; (ii) マーケティングの統計分析事例：市場機会の発見（因子分析）、製品開発（コンジョイント分析）、新製品の普及（Bass モデル）; (iii) 現代マーケティングと個別化対応; (iv) 異質性モデリングとベイズ統計; (v) 応用事例：価格閾値の推定と個別化戦略の可能性.

* 本稿は 2025 年 4 月 28 日に行われた統計数理研究所・統計エキスパート人材育成プロジェクト・連続講義シリーズ：社会の中の統計科学 第 3 回「マーケティングと統計科学」の講義ノートに基いている。

† 東京理科大学経営学部ビジネスエコノミクス学科教授

1　現代マーケティングの社会的背景と諸問題

近年、ネットワーク技術や IoT（Internet of Things）の急速な進展に伴い、社会は「データの世紀」とも呼ばれる時代に突入している。膨大なデータがリアルタイムかつ自動的に生成・蓄積されており、その量は指数関数的に増大している。かつては、こうしたデータの蓄積に伴う記録容量や保存コストが問題視されていたが、近年ではむしろ、それらの処理や分析にかかる計算資源やエネルギーの制約が新たな課題として浮上している。このような環境変化を受けて、我が国では「Society 5.0」および「超スマート社会」の実現を掲げ、データを活用した社会構造の変革を推進している。Society 5.0 とは、第 4 次産業革命の技術革新を背景に、主観的な判断や経験則に依存した意思決定から、客観的データに基づいた意思決定（エビデンス・ベースド・デシジョンメイキング）への移行を目指す構想である。

この構想の根底には、経済の「サービス化」がある。日本を含む先進諸国では、GDP の約 70%、および就業者の 70% 以上がサービス産業に従事しており、従来の製造業中心の経済構造から、サービス主導型の構造へと移行している。このサービス分野におけるイノベーションを推進するためには、無意識的に蓄積された多種多様なビッグデータを有効活用することが不可欠とされている。とりわけ、雑多で非構造的なデータ群を横断的に活用し、「必要なものを、必要な人に、必要なときに、必要なだけ」提供するという構想は、マーケティング分野において近年注目を集めている「パーソナライゼーション（個別化対応）と深く通底する概念である。1990 年代以降、マーケティング領域では、こうした個別対応型のアプローチを実現するためのデータ駆動型手法の研究が活発に進められてきており、実務面でも徐々にその実装が進んでいる。

上で述べたような「必要なものを、必要な人に、必要なときに、必要なだけ提供する」という構想は、マーケティングにおいては「個別化対応（Personalization）」として実装されている概念である。これは、単に製品やサービスを提供するだけでなく、顧客のニーズや文脈に即した対応を行うことで、より高い満足度やロイヤルティの獲得を目指すアプローチである。この個別化対応の発想は、1990 年代からすでにマーケティング分野において研究が始まっており、とくにデジタル化の進展に伴って実用的な応用が拡大してきた。デジタルデバイスの普及とオンライン行動のトラッキング技術の発展により、企業は消費者ごとの購買履歴、閲覧傾向、接触メディアなど多様なデータをリアルタイムで収集可能となった。このようなパーソナライゼーションは、単なるセグメンテーション（市場細分化）を超え、One-to-One Marketing とも呼ばれる、一人一人に最適化されたコミュニケーションや提案を実現するものとして位置づけられている。

また、現代においては、機械学習やベイズ統計を用いた「異質性モデリング」によって、消費者個人ごとの反応性や選好の推定が可能となり、理論と実務の両面でその精度が飛躍的に向上している。マーケティングにおけるこのような個別最適化の動向は、Society 5.0 が目指す社会的課題解決型イノベーションとも軌を一にする。すなわち、情報技術とデータを活用することによって、社会の多様なニーズにきめ細かく応えうる社会基盤の形成が、経済および産業の発展と共に進められている。

2　マーケティングとは何か

2.1　制御変数としての 4P

マーケティングとは何かを明確に定義することは、その実践と分析において不可欠である。本稿では、理工系を含む広範な分野の読者にも理解可能なように、マーケティングを数理的な最適化問題として定式化する。すなわち、マーケティングとは「企業が制御可能な複数の戦略的変数を用いて、売上、利益、市場占有率といった目的関数を最大化もしくは最適化しようとする活動」であると定義できる。この戦略的変数の代表格が「4P」として知られるフレームワークである。4P とは以下の通りである。

(i) Product（製品）：企業が提供する製品やサービスの内容や仕様
(ii) Price（価格）：設定される価格水準や価格戦略
(iii) Promotion（販売促進）：広告、キャンペーン、クーポン等の販売刺激活動
(iv) Place（流通）：製品・サービスをどのように消費者の手元に届けるかに関する流通戦略

これら 4 つの要素は企業が自律的に設計・変更可能な制御変数として機能し、それらを駆使することで企業目標の最適化を図る。ただし、実際の意思決定には市場環境、競合の動向、景気、規制といった外部的な制約条件が付随するため、制約付き最適化問題として捉えるのが適切である。表 2.1 は、マーケティング活動のライフサイクル全体を時間軸に沿って整理したものである。ここでは、製品・サービスの企画・開発段階から始まり、市場導入、成長、成熟、撤退に至るまでの各フェーズで、どのような分析課題が存在し、それにどのような統計的手法が対応するかを視覚的に整理している。

表 1: マーケティング活動の段階と分析手法の対応関係（例）

時間軸フェーズ	主な課題	対応する統計的手法
市場機会の探索	ニーズの特定、競合分析	因子分析、セグメンテーション、知覚マップ
製品開発	属性評価、製品仕様設計	コンジョイント分析
市場導入と初期販売	製品の普及予測	バスモデルなどの普及モデル
成熟・競争環境の管理	ライフサイクル戦略	クラスタリング、ポジショニング分析

2.2　範囲と段階：市場発見からライフサイクルマネジメント

図 2.1 左は著者によるマーケティングのテキスト「現代マーケティングリサーチ（有斐閣）」の目次を示している。さらに右図は各章が扱う様々な統計分析手法のマーケティングにおける位置づけを表している。

すなわちマーケティング活動の出発点は、対象とする市場においてどのような「機会」が存在するのかを発見することである。これを「市場機会の発見」と呼び、この段階では知覚マップやセグメンテーション分析、同時マップ分析などの手法が活用される。これにより、消費者の認知空間における競合との位置関係が可視化

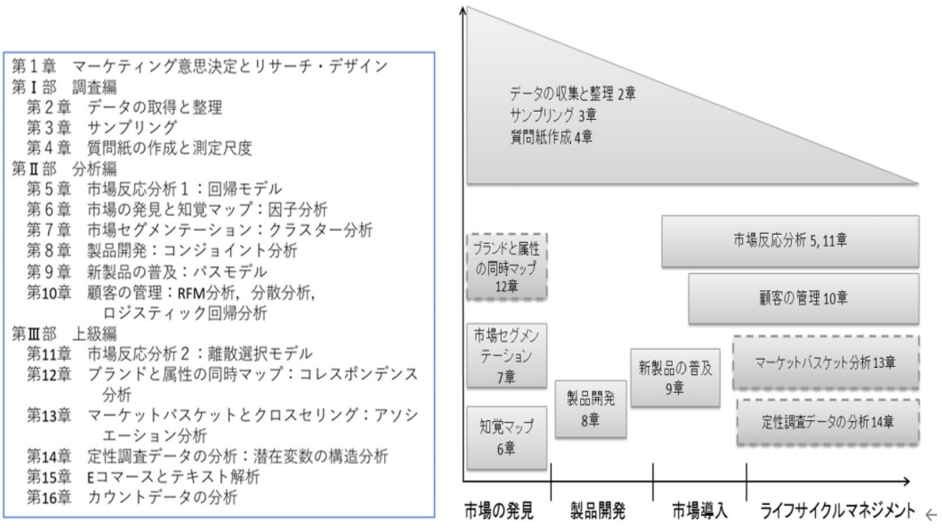


図 2.1: マーケティングの統計分析と戦略上の位置づけ（照井・佐藤 (2022)）

され、市場における未充足領域の存在が明らかになる。次に、その分析結果に基づいて、具体的な製品開発フェーズへと移行する。この段階では、消費者が製品の各属性に対してどのような選好や効用を持っているのかを明らかにするために、コンジョイント分析などの統計手法が用いられる。製品が市場に導入されると、初期販売データから将来の販売量を予測する必要が生じる。特に製品ライフサイクルの初期段階では、限られたデータから普及曲線を推定する必要があり、Bass モデルなどの製品普及モデルがその目的に適している。その後、製品や市場が成熟するにつれて、競争環境や消費者嗜好の変化に対応した「ライフサイクル・マネジメント」が求められるようになる。この段階では、市場全体を複数のセグメントに分割し、それぞれに異なる戦略を適用する必要があるため、クラスタリングやポジショニング分析といった多変量手法が再び重要性を持つ。

このように、マーケティングとは単一の局面に限定される活動ではなく、製品の一生を通して多段階かつ多次元にわたる統計的・戦略的アプローチを必要とする動的なプロセスである。

3　マーケティングの統計分析事例: (I) 市場機会の発見

3.1　因子分析を用いた市場構造分析

本節では、消費者の知覚に基づく市場分析手法の一つとして、因子分析を利用したアプローチを説明する。因子分析は、多変量解析のスタンダードな手法であり、複数の観測変数を少数の潜在的因子に集約することができる。消費者のブランドイメージに関する認知構造を明確にするために、この手法は非常に有効である。因子分析の基盤として、例えば教育における科目の成績に関する因子分析を考えることができる。具体的には、文系と理系という二つの大きな因子が存在すると仮定した場合、国語、英語、社会科などの文系科目と、数学、理科といった理系科目に分けられる。各科目は、それぞれ異なる因子負荷量を持ち、例えば文系科目は文系因子に強く関連し、理系科目は理系因子に強く関連することが予想される。

この因子分析を消費者のブランド認知に適用する場合、ブランドに関する複数の要因（例：パッケージデザ

表 3.1 ブランド評価データ (6 章: 照井・佐藤 (2022)))

製品名	味	パッケージ	広告宣伝	素材栄養素	キャンペーン
1. ハッピーターン	89	63	32	51	37
2. 雪の宿	73	46	25	48	32
3. ぽたぽた焼き	65	45	17	32	21
4. 黒豆せんべい	72	33	2	50	9
5. まがりせんべい	70	29	11	35	21
6. チーズアーモンド	71	20	10	38	24
7. 手塩屋	71	38	3	38	13
8. ばかうけ	48	39	10	16	16
9. 粒より小餅	49	26	17	27	30
10. 田舎おかき	51	16	1	38	5
11. うまい！堅焼き	48	15	4	36	1

出所：日経 MJ 新聞 2011.2.13

イン、広告宣伝、キャンペーンイベントなど）について調査を行うことができる。例えば、日経流通新聞が実施した調査において、11 種類のせんべいを対象に、味やパッケージデザイン、広告宣伝などの要因について消費者に評価を求めた。調査結果は、これらの要因が高い相関を持つことが示され、因子分析を用いて二つの主要な因子に集約されることとなった。

因子スコアを計算した後、それを二次元空間にマッピングすることで、市場におけるブランド間の位置関係を視覚化することができる。このマッピング手法は「知覚マップ」と呼ばれ、ブランド間の競争関係を分析する上で非常に有用である。例えば、マーケティング因子（キャンペーンや広告宣伝など）と製品因子（実際の商品品質やデザインなど）の二軸でマッピングを行い、それぞれのブランドの因子スコアをプロットする。

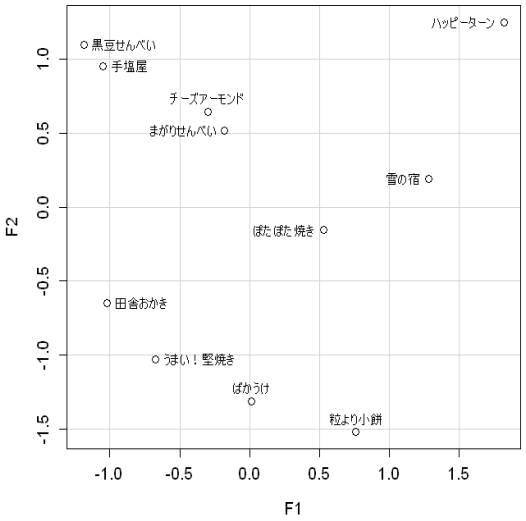


図 3.1: 知覚マップ: 6 章照井・佐藤 (2022))

知覚マップを通じて、消費者にとって類似したブランドが近接して配置されることが分かる。このため、近接しているブランドは競合関係にあると解釈され、逆に隙間が広がっている場所は市場における新たな機会を示唆していると考えられる。このようにして、競合の少ない市場セグメントを特定し、マーケティング戦略を最適化するための指針を得ることができる。

また、サブマーケットの特定に関しては、因子分析によって得られた結果を基にクラスタリング手法を適用

することが一般的である。クラスター分析では、因子スコアを基に市場を複数のグループに分け、それぞれのグループが持つ特性を把握する。このような分析により、競合が集中している市場セグメントと、競争が少ない市場セグメントを明確に区別することができる。

3.2 マーケティングの統計分析事例: (II) 製品開発

3.2.1 コンジョイント分析における属性と水準の設定

コンジョイント分析を適用する際、最初に行うべきは製品やサービスの「属性」と「水準」の定義である。これらは消費者が製品を評価する基準となり、効用を算出するための基本的な構成要素を形成する。属性は製品の特徴を示し、各属性にはそれぞれ異なる水準が設定される。たとえば、電化製品においては、「バッテリー持続時間」、「保証期間」、「色」といった属性が考えられる。それぞれの属性に対して、適切な水準を設定することが重要であり、これによって消費者が各製品をどのように評価するかが決まる。「属性の例とその水準」例えば、電化製品を対象とした場合、以下のような属性とその水準を考えることができる。

- バッテリー持続時間: 3 時間、6 時間、9 時間
- 保証期間: 1 年、2 年
- 色: 赤、シルバー、黒

これらの属性と水準を組み合わせることで、さまざまな製品の選択肢が生まれる。それぞれの製品には異なる属性水準の組み合わせが設定され、消費者はその組み合わせに基づいて製品を評価することとなる。重要なのは、これらの組み合わせが消費者の選好にどのように影響するかを評価することにある。

表 2: 属性と水準, 部分効用 (8 章 照井・佐藤 (2022))

属性	水準			部分効用		
バッテリ時間	6 時間	4 時間	2 時間	u_{11}	u_{12}	u_{13}
保証期間	2 年	1 年	-	u_{21}	u_{22}	-
色	赤	シルバー	黒	u_{31}	u_{32}	u_{33}

3.2.2 効用の推定方法

コンジョイント分析では、消費者が製品の各属性にどれだけの価値を置いているかを定量的に測定する。各属性に対応する効用（「部分効用」）を算出し、その合計が「全体効用」となる。例えば、消費者が「バッテリー持続時間が 6 時間」、「保証期間が 2 年」、「色が赤」の製品を選んだ場合、その効用は各属性の部分効用を足し合わせることによって求められる。このようにして、消費者の選好が属性ごとにどれだけ重要視されているかを明確にすることができる。

コンジョイント分析における効用の推定は、選好順位からつくられる目的変数 y を属性水準の有無を表す複数のダミー説明変数で説明する重回帰分析を用いて行われる。この分析では、製品の属性水準をダミー変数と

して表現し、これらを独立変数として回帰分析を実施する。回帰分析の結果として得られる回帰係数が、それぞれの属性水準に対応する部分効用となる。例えば、バッテリー持続時間が 3 時間、6 時間、9 時間の 3 水準を持つ場合、それぞれの水準に対応するダミー変数を作成する。次に、保証期間が 1 年と 2 年の 2 水準を持つ場合、1 つのダミー変数が設定される。同様に、色に関しては 3 水準（赤、シルバー、黒）に対応する 2 つのダミー変数が設定される。このようにして、各属性水準に対応するダミー変数を使って回帰モデルを構築し、消費者の選好を数量的に表現する。

データ収集のためには、実際に消費者に対して製品の選好順位を尋ねる調査を実施することが一般的である。たとえば、18 種類の異なる製品を設定し、それぞれの製品に対して消費者に好ましい順番を付けてもらう。これにより、消費者の選好を数値化することができ、そのデータを基に回帰分析を行い、効用の推定を行う。

3.3 効用測定モデル

まず、説明変数が属性水準によってカテゴリーに分けられるため、0, 1 のダミー変数で定義される重回帰モデルにより測定できる。上述の電化製品の新製品開発の場合、2 つの属性で 3 水準、一つの属性で 2 水準あるので、可能性のある新製品の組み合わせは全部で 18 通りある。

このとき、コンジョイント分析では、３つの属性の部分効用を重回帰モデルの説明変数の一部分に対応させた形で表現した効用測定モデルを設定する。すなわち、重回帰モデル

$$\begin{aligned} y = \alpha + \beta_{11}x_1 + \beta_{12}x_2 + \beta_{21}x_3 \\ +\beta_{31}x_4 + \beta_{32}x_5 + \varepsilon \end{aligned} \tag{3.1}$$

を設定して、回帰係数の推定値 $\hat{\beta}_{32}$ から部分効用の値を求める。

ここで部分効用を定義する 5 つの説明変数は実数値ではなく、以下で説明する水準の違いを表わすカテゴリ変数である。この属性水準を区別するカテゴリ説明変数は、バッテリ時間の属性については 3 水準（カテゴリ）であるので２つの説明変数を導入し以下のように定義する。

$$x_1 : \text{時間 } 1 = \begin{cases} 1, & 6 \text{ 時間} \\ 0, & \text{それ以外} \end{cases} \quad ; \quad x_2 : \text{時間 } 2 = \begin{cases} 1, & 4 \text{ 時間} \\ 0, & \text{それ以外} \end{cases} \tag{3.2}$$

保証期間については 2 水準であるので

$$x_3 : \text{保証} = \begin{cases} 1, & \text{保証 } 2 \text{ 年} \\ 0, & \text{それ以外} \end{cases} \tag{3.3}$$

と定義され、色については 3 水準で

$$x_4 : \text{色 } 1 = \begin{cases} 1, & \text{赤} \\ 0, & \text{それ以外} \end{cases} \quad ; \quad x_5 : \text{色 } 2 = \begin{cases} 1, & \text{シルバー} \\ 0, & \text{それ以外} \end{cases} \tag{3.4}$$

のように定義する。

また、選択ベースコンジョイント（CBC）分析もデータ収集手法として有効であり、消費者に対していくつ

かの製品の選択肢から最適な製品を選ばせる方法である。この方法により、消費者が実際の選択状況においてどのように意思決定を行っているかを反映させることができる。

3.3.1 結果の可視化と解釈

得られた回帰係数を基に、各属性の部分効用を可視化することで、消費者がどの属性を重視しているかを明確にすることができる。例えば、保証期間が 2 年であることが非常に重要であり、色が赤であることも重要な要素であることが分かる。バッテリー持続時間についても、長時間の方が消費者に好まれる傾向がある。このような分析結果を基に、製品設計やマーケティング戦略を策定することが可能となる。

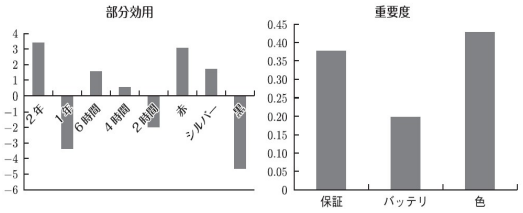


図 3.2: 部分効用と属性の重要度 (8 章：照井・佐藤,2022)

3.4 マーケティングの統計分析事例: (III) 製品の普及プロセスの Bass モデル

3.4.1 新製品の分類と普及過程

新製品の開発にはさまざまな目的が存在し、それに応じて製品の普及戦略も異なる。新製品は、既存の製品カテゴリーに新たに加わるライン拡張、新しいブランド名で製品を投入するマルチブランド戦略、全く新しいカテゴリーに製品を投入する新ブランド戦略に分類される。Bass モデルは、これらの戦略における製品の普及パターンを予測するために有効であり、それぞれの戦略に応じた販売予測を立てることができる。製品の普及過程は、以下の 4 つの段階に分類される。

(1) 導入期：新製品が市場に登場する段階。
(2) 成長期：製品が急速に普及し、売上が増加する段階。
(3) 成熟期：市場での競争が激化し、売上の伸びが鈍化する段階。
(4) 衰退期：製品の需要が減少し、最終的に市場から撤退する段階。

これらの段階を理解することで、企業は適切なタイミングで販売戦略を調整することができる。製品の普及は、企業のマーケティング戦略において重要な要素であり、特に新製品の市場投入時には、初期の売上データをもとに今後の販売推移を予測することが求められる。こうした予測を行うために、Bass モデル（Bass, 1968）は有効な統計的手法として広く用いられている。本稿では、バスモデルの概要とその適用方法について、特に製品普及の過程を予測するためのモデルとしての有用性を論じる。

3.4.2 Bass モデルの概要と数理的基礎

Bass モデルは 1969 年に Frank Bass によって提案された。(Bass, 1969) このモデルは、新製品が市場に登場した際、消費者の購買行動を 2 つの主要なグループに分類し、それに基づいて製品の普及過程を予測するものである。Bass モデルでは、消費者はイノベーター（革新者）と模倣者の 2 つのカテゴリーに分けられる。イノベーターは新しい製品を他者の影響を受けずに自主的に採用し、模倣者は市場の普及状況を見てから製品を購入する。バスモデルは、製品の普及過程を予測するための数学的手法として、微分方程式に基づいて記述されるが、消費者の購買行動を記述するための基本的な数式は以下のように表される。

$$f(t) = \frac{dF(t)}{t} = [p + qF(t)] \tag{3.5}$$

ここで $f(t)$ は t 時点の普及率 (採用率), $F(t)$ は t 時点の累世普及率（累積採用率）, p, q はそれぞれイノベーターと模倣者の購買比率を示すパラメータである。この式は、製品の普及が時間とともにどのように進行するかを示している。また

$$\frac{f(t)}{1 - F(t)} = [p + qF(t)] \tag{3.6}$$

と変化したとき、左辺は t 期までに購入していないことを条件に、t 期に購入する確率（ハザード）を表し、これが右辺の革新者 p と模倣者 $qF(t)$ の和で表されることを意味し、t 期までに購入しないことを条件として、t 期に購入する確率が $\frac{f(t)}{(1-F(t)}$ であり、それがイノベーター p と模倣者 $qF(t)$ の和で表されるとするモデルである。

この微分方程式モデルを単純に離散化して計量モデルを構成する方法を紹介する。この計量モデル化の手順は図 xx で説明されている。ここで m：潜在市場規模（最大販売可能量)、n_t: t 期における購入者数、N_t: t 期

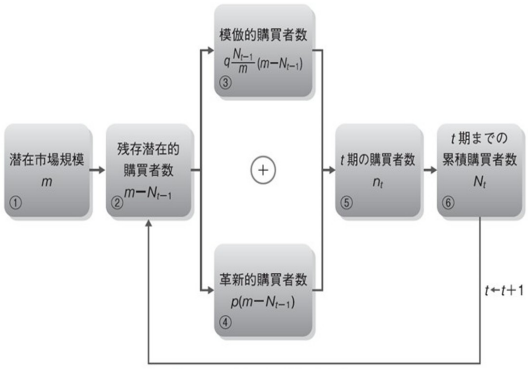


図 3.3: バスモデルの計算手順 (9 章：照井・佐藤,2022)

までの累積購入者を意味している。この手順によって計算される t 期の新規購入者数は図 3.5 第 1 式にあるように、N_{t-1} および N_{t-1}^2 の 2 つの説明変数をもつ重回帰モデルとしてパラメータが推定できることを示している。Bass モデルを適用するには、まず最初の数週間の売上数と累積購入者数のデータを収集し、これらを従属変数および説明変数とする回帰分析により回帰係数パラメーター (a, b, c) の推定値を求め、それを逆算して

パラメータ (p, q, m) を求める。

すなわち

$$n_t = pm + (q - p)N_{t-1} - \frac{q}{m}N_{t-1}^2 \tag{3.7}$$

ここで $a = pm, b = (q - p), c = \frac{q}{m}$ としてパラメータ変換すると

$$n_t = a + bN_{t-1} - cN_{t-1}^2 \tag{3.8}$$

と表現できる。最後に $m = \frac{-b-\sqrt{b^2-4ac}2c}{,} p = \frac{a}{m}, q = p + b$ と逆算する。

3.4.3 モデルの結果とその解釈

Bass モデルによって得られた推定結果は、製品の普及パターンを示す。特に、イノベーターと模倣者の購買比率 (p, q) が異なる場合、普及の速度やパターンが大きく変化することがわかる。例えば、p が大きい場合、革新的な消費者が早期に購入し、その後急激に普及が進むことが予想される。一方、q が大きい場合、普及は緩やかに進行し、後期多数採用者の影響が大きくなる。

Bass モデルは、新製品の市場投入における消費者の購買行動を予測するための強力なツールである。特に、初期の販売データを基にして、製品の普及過程を予測することができ、企業にとって重要なマーケティング戦略をサポートする。モデルの精度を高めるためには、実際の市場においてはモデルを適切に調整することが重要である。

4 現代のマーケティング：個別対応

4.1 マーケティングの考え方の変遷

近年のマーケティング研究および実務において、個々の消費者のニーズや行動特性に応じた個別対応（パーソナライゼーション）の重要性が強く認識されている。こうした潮流は、Society 5.0 を背景とした「必要なときに、必要な人に、必要なものを届ける」といった価値創出の枠組みと深く結びついており、従来の均質的なマーケティング・アプローチとは一線を画すものである。本節では、マーケティング理論の歴史的展開を概観した上で、現在求められている個別化アプローチおよび統計的モデリングによる異質性の捉え方について論じる。

マーケティングの初期段階においては、消費者を一様な存在と見なし、全体に向けて同一のメッセージや製品を届けるマスマーケティングが主流であった。この枠組みにおいては、いわゆる「代表的な消費者」を想定し、その嗜好や行動に基づいて市場を理解しようとする立場が支配的であった。しかし、現実の市場における消費者は、年齢、性別、所得、価値観など多様な側面において異質であり、単一の代表像に還元することは不可能であるという認識が次第に広まった。このような背景のもとで登場したのが市場細分化（セグメンテーション）の概念である。これは、消費者をいくつかのグループ（セグメント）に分類し、それぞれに異なるマーケティング戦略を展開することを目的とするものである。このセグメンテーションは、たとえば「男性と女性」

「子どもと大人」といったデモグラフィック属性によって市場を区分することにより、よりターゲットに即したアプローチを可能とした。しかし、それでもなお「セグメント内の異質性」という問題は未解決であり、より粒度の細かい対応が求められるようになっていった。

セグメンテーションの論理をさらに発展させたものが、ターゲティング戦略および One to One マーケティングである。このアプローチでは、単にグループごとに異なる対応をするのではなく、個々の消費者の属性、履歴、嗜好、行動に基づいて、最適な対応をリアルタイムで行うことを目的としている。このような精緻なマーケティングが実現可能になった背景には、ビッグデータおよび ICT 技術の進展がある。消費者の購買履歴、Web 行動、位置情報など多様なデータを統合的に分析することで、従来では不可能であった「一人ひとりの消費者理解」が現実となっている。

4.2 個別対応マーケティング

個別対応を理論的に支えるためには、消費者の異質性（heterogeneity）を数量的に捉える必要がある。そのための代表的な方法論が、階層ベイズモデル（hierarchical Bayesian modeling）である。この手法は、個人レベルのパラメータ（例：価格感度、ブランド選好など）が集団内でどのように分布しているかを、ベイズ推論によって推定する。階層ベイズモデルの利点は、すべての消費者に対して「固有のパラメータ」を割り当てることができる点にあり、個人ベースのマーケティング施策と極めて親和性が高い。たとえば、購買履歴に基づく個別価格設定（personalized pricing）やレコメンドシステムなどは、まさにこの種の統計モデリングの成果と言える。

マーケティングの発展は、消費者の理解を「平均的な代表者像」から「個々のパーソナリティ」へと深化させてきた。その背景には、社会的要請としての多様性尊重と、技術的裏付けとしてのデータ科学の進歩がある。特に現代においては、階層ベイズモデルに代表される異質性モデリングが、個別化対応を支える理論的かつ実務的な基盤となっている。

4.3 ベイズ統計の枠組みと特徴

統計学における 2 大潮流の一つであるベイズ統計は、頻度論（フィッシャー流）と対照的な立場を取る推論体系である。頻度論において未知パラメータは固定された定数と見なされるのに対し、ベイズ統計では未知パラメータを確率変数と捉え、事前分布を与えることで不確実性を表現する。本稿では、ベイズ統計の理論的枠組みとその実用的意義について概観し、特に少数データや複雑モデルにおける推論の可能性について論じる。

ベイズ統計の中心的な考え方は、事前分布とデータから得られる尤度を統合し、事後分布を更新・構築することで推論を行う点にある。このプロセスはベイズの定理によって形式化される。すなわち、

$$p(\theta|y) = \frac{p(\theta)(y|\theta)}{p(y)} \tag{4.1}$$

により、観測データ y を得た後のパラメータ θ の分布（事後分布）を計算する。この枠組みにより、統計解析

において既存の知識や理論、経験を事前分布として取り入れることが可能となる。

頻度論的アプローチでは、標本を無限に繰り返し取得可能であるという前提のもと、点推定・区間推定・仮説検定が中心的手法となる。これは中心極限定理に支えられており、漸近的に正規分布へと収束することを利用する。一方、ベイズ統計では有限のデータであっても、事前情報と統合することにより信頼できる推論が可能となる。ベイズ推論の実装において、積分評価が計算上のボトルネックとなる。1990 年以前は、解析的に積分可能な限られた事前分布（例：共役事前分布）に依存する必要があった。しかし 1990 年代以降、マルコフ連鎖モンテカルロ法（Markov Chain Monte Carlo: MCMC）の発展により、数値的に任意の積分が近似可能となった。これにより、ベイズ統計は理論的な枠組みから実用的ツールへと大きく飛躍した。

(1) 少数データとベイズ推論ベイズ統計は、少数データにおける推論にも強みを持つ。データが少ない場合でも、妥当な事前分布を設定することで精度の高い推定が可能である。これは、問題の背後にある理論的知見や実務的経験を統計モデルに反映することができるためである。たとえば、消費関数モデルにおいて、標本数が 10 と少ない状況でも、事前分布を活用すれば限界消費性向といったパラメータの推定が可能となる。
(2) 複雑なモデルへの対応ベイズ統計は、非正則な問題や構造変化モデルにも柔軟に対応できる。たとえば、変化点回帰モデルにおいて、x の値によって回帰係数が急激に変化するような状況がある。頻度論的手法では、このようなモデルの推定は困難であるが、ベイズ統計では階層モデルなどを通じて自然に推定が可能である。
(3) 異質性モデリング個別対応型モデル、すなわち各個体が異なる回帰係数を持つようなモデルにおいても、ベイズ階層モデルを用いることで、個体間の共通性（例えば正規分布の平均・分散など）を仮定しつつ、個体ごとのパラメータを推定することが可能である。このようなアプローチは、データが極めて少ない個体に対しても有効である。

4.4 異質性モデリングにおける統計的アプローチ：階層ベイズモデル

4.4.1 階層ベイズモデルの基本構造

消費者行動の分析において、各個人は異なる特性や行動様式を示す一方で、一定の共通性も持ち合わせている。このような「異質性と共通性の同時モデリング」は、マーケティングや行動科学などの分野で重要な課題となっている。階層ベイズ（Hierarchical Bayes, HB）モデルは、このような問題設定に対して柔軟かつ強力な推定手法を提供するものである。本節では、HB モデルの基本的な考え方と構造について説明する。

階層ベイズモデルでは、各消費者に固有のパラメータベクトルを設定し、それらが共通の分布に従うと仮定する。この分布のパラメータは、全体の消費者集団に関する事前的な知見や属性データ（例：年齢、性別、家族構成）をもとにモデル化される。図 4.2 に示すようにもまた上位の事前分布を持つことで、階層構造を形成する。

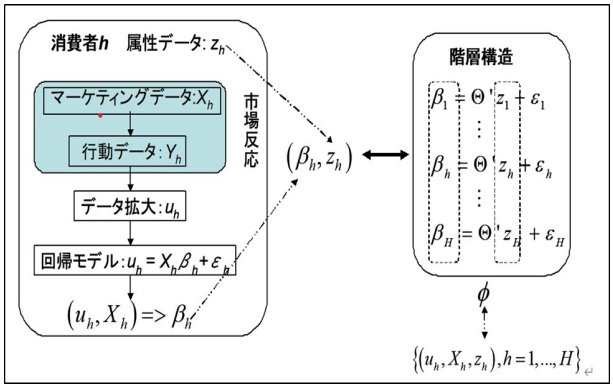


図 4.1: 階層ベイズ離散選択モデル

- 個人レベルのモ行動デル（レベル 1）：ここで、 Y_h は消費者 h の観測データ、X_h は対応する説明変数,β_h は個別の異質性パラメータである。
- 個人パラメータの事前分布（レベル 2）：z_h は消費者属性、θ は共通な回帰係数であり、これにより異質性と共通性のトレードオフを統計的に制御する。

4.4.2 直観的説明

図 4.1 は、各消費者の行動パターン（顔の特徴）を例にとり、個別データのみでは十分な情報が得られない状況において、集団全体の情報（共通の顔）を活用することで推定精度を高めるという HB モデルの直感を示している。個人の観測数が限られる場合でも、共通構造を利用することで、より安定した個別推定が可能となる。

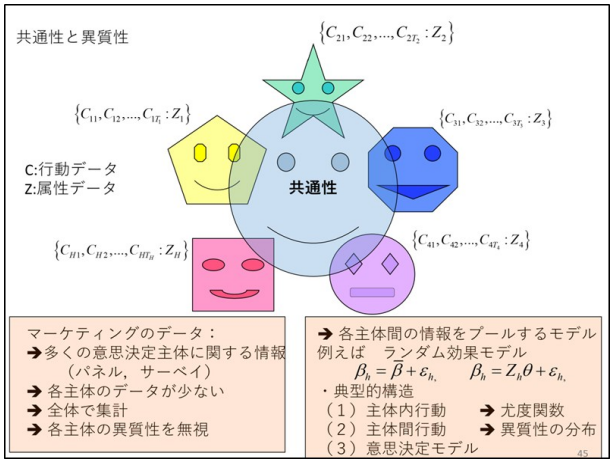


図 4.2: 異質性モデリングと階層モデル

4.4.3 モデル推定と MCMC

HB モデルの推定には、MCMC（マルコフ連鎖モンテカルロ）法が広く用いられている。これは事後分布からのサンプリングを通じて、複雑なモデル構造にもかかわらず推定可能とする強力な数値計算手法である。特

に、MCMC により得られる事後サンプルからは、予測分布の生成や信頼区間の評価なども容易に行える。

階層ベイズを含むベイズ推定の理論的正当性は、情報理論の観点からも支持されている。赤池（Akaike, 1980）は、ベイズ推定がエントロピー最小化、すなわち情報損失の最小化という意味で最適であることを理論的に証明している。この視点は、モデル選択や予測精度の観点からベイズ手法を用いることの意義をより一層明確にする。階層ベイズモデルは、消費者異質性のモデリングにおいて、データの希薄性やモデルの複雑性に対処するための有力な枠組みである。個別と共通の情報をバランスよく活用することで、より実用的かつ理論的に整合的な推定を実現する。

4.5 階層ベイズモデルを用いたマーケティングデータ分析の研究例

4.5.1 マーケティングデータの特徴と課題

本節では、階層ベイズモデルを活用したマーケティングにおける応用例について、研究レベルの分析を紹介する。特に ID 付き POS（Point of Sale）データを対象とし、消費者の異質性を捉える方法論について検討する。

まずマーケティングデータには多様な形式があるが、近年ではスーパーマーケット等の POS システムにより収集される ID 付きスキャナーデータが広く用いられている。このデータは、会員登録時に得られる属性情報（年齢、性別、家族構成等）と、実際の購買履歴（購入商品、価格、プロモーション状況等）から構成される。このようなデータには以下のような特徴と課題が存在する：

- 情報の不均一性：全体の消費者数は多い一方で、個々の消費者ごとの購買回数は少ない場合が多い。
- 離散データの性質：選択されたブランド等は離散的に記録される。

4.5.2 分析方法：階層ベイズモデルの適用

上記の課題に対応するため、階層ベイズモデルが有効である。このモデルは、消費者ごとのパラメータが、全体の平均を中心とした分布から生起すると仮定する。ここで、z_h は消費者 h の属性データを示し、θ は回帰係数ベクトルである。このように、β_h の異質性を属性情報で説明しつつ、共通部分を全体でプールされた事前分布として扱う。

離散選択データに対しては、データオーグメンテーションにより連続的な潜在効用変数を導入し、構造の推定を可能にする。具体的には、選択されたブランドに対応する効用が最も高かったと仮定し、その効用に正規分布等の連続分布を仮定することで、MCMC（Markov Chain Monte Carlo）によるベイズ推定が行える。(Allenby and Rossi, 1996; Rossi et al. 2005)

詳細は割愛するが、モデルは次の複数の選択肢の中から一つを選択する離散選択モデルを尤度関数として用いる。

1. 消費者行動データ＝＞選択肢の中から 1 つを選択 (A, B C から B を選択) ＝＞離散選択モデル（ロジッ

トモデル、プロビットモデル）＝＞尤度関数を構成

2. 消費者属性データ＝＞年齢、性別、家族人数、職業 e.t.c. ＝＞消費者異質性の階層モデル

4.5.3 分析事例：価格閾値の推定と価格の個別化戦略の可能性

分析対象とするスキャナーデータは、以下のような情報を含む：

- 購買日と回数（例：4 回）
- 商品カテゴリー（例：A, B, C）
- 各回の選択ブランド
- 各ブランドの価格とプロモーション有無

この情報を用いて、消費者ごとの価格感度やプロモーションへの反応度を推定する。具体的には、離散選択モデル（多項ロジット／プロビット）から定義される尤度関数と事前分布としての異質性モデルと組み合わせて階層ベイズ的に推定する。

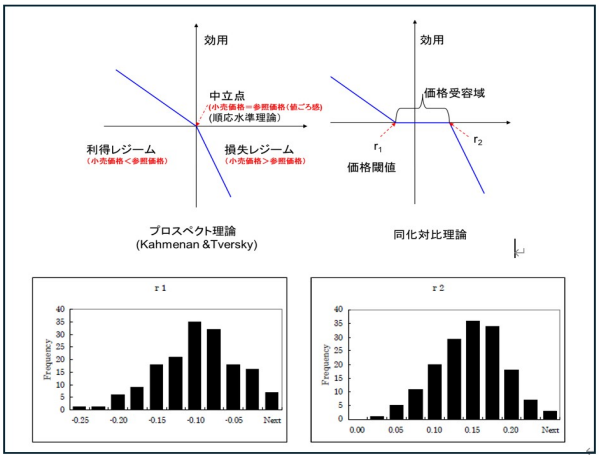


図 4.3: 価格閾値モデルと個別推定の結果

消費者が商品を購入する際には、頭の中で形成された参照価格と実際の小売価格を比較して損得を判断するという理論（Prospect Theory）を前提とする。このとき、参照価格と実際価格の差が一定範囲（価格受容域）内であれば、消費者は価格の変化に対して鈍感に反応する。この価格受容域は個人ごとに異なるため、階層モデルにより個別に推定することが可能である。消費者パネルデータを用いてこれを推定した結果（Terui and Dahana, 2006a, b）では、次の結論が得られた。

- ディスカウント（利得）側の閾値は、25% 以上値引きしないと反応しない消費者もいれば、0.5% でも反応する者も存在する。
- 値上げ（損失）側の閾値についても同様であり、価格戦略の立案に有用な知見が得られる。

このような階層ベイズ的アプローチを取ることで、マクロ的な市場分析とミクロ的な消費者行動の理解を統

合することができる。個別価格設定が可能なオンライン環境では、こうした推定結果に基づく価格の個別化が現実的に実施可能となっており、今後のマーケティング研究と実務への応用が期待される。

5 おわりに

マーケティングにおける統計科学の活用は、古典的な解析手法から、現代のベイズ的アプローチや機械学習を用いた手法にまで、大きく広がりつつある。とりわけ、個別対応が求められる現代の消費者行動を分析するうえで、階層ベイズモデルや MCMC（マルコフ連鎖モンテカルロ法）を用いた手法は、理論的にも実務的にも極めて有力なツールとなっている。今後は、データ科学と統計科学の融合により、マーケティング研究の一層の深化が期待される。さらに、SNS をはじめとした消費者間あるいは消費者と企業間のコミュニケーションにより生成される、大規模かつ非構造的なデータも、マーケティングにおいて極めて重要な情報源である。本講義ではこれら非構造データの分析は時間の制約上割愛した。

これらのデータに対しては、自然言語処理や機械学習を用いた分析が進められている。しかし現状では、主に予測性能の向上に焦点が当てられており、識別性のない特異なモデルを扱うことも多いため、モデルの結果を解釈して戦略策定へとつなげるには課題が残されている。ビジネスにおける計画策定を可能にするためには、因果関係に基づく計量モデル、すなわち $Y = f(X)$ の形をとる戦略モデルが求められる。しかしながら、ビジネスのステークホルダーである人間は、個性や行動の揺らぎが大きく、また自然現象のように安定した「物理定数」に相当するものを社会経済システムに想定することは困難である。そのため、問題の背景にある分野知識や理論を総動員し、有用な情報を抽出する仕組みが必要となる。こうした場面では、先験的知識としての経験・理論・専門的合意（すなわち事前分布）と、観測された経験データ（尤度）を統合することにより、複雑な事象を事後分布というかたちで認識し、情報を更新していくことが有効である。これにより、検証・予測・制御といった意思決定への応用が可能となる。

このような背景のもと、ビジネスデータの分析にはベイズ統計に基づく統計モデリングが極めて有効である。とりわけ、以下のような目的に対してベイズ統計は重要な基盤技術となる：

(1) 個性の推定に基づく個別対応
(2) 次元圧縮を通じた大規模データへの対応
(3) 結果から原因を推論する逆問題へのアプローチ
(4) 異なる情報を統合することによる新たな知の創出

詳細は、自身の研究を紹介しながら、これらの視点に基づいてビジネスデータの分析を展望している照井 (2024)、具体的な AI/機械学習手法の入門的解説は照井 (2018) を参照されたい。

参考文献

1. 照井伸彦, 佐藤忠彦著（2022)「(新版）現代マーケティング・リサーチ―市場を読み解くデータ分析」, 有斐閣.
2. 照井伸彦 (2010)「R によるベイズ統計分析」, 朝倉書店
3. 照井伸彦 (2008)「ベイズモデリングによるマーケティング分析」東京電機大学出版局
4. 照井伸彦 (2018)「ビッグデータ統計解析入門－経済学部で学ばない統計学」日本評論社.
5. 照井伸彦 (2008b), 「価格閾値の推定と価格カスタマイゼーションの可能性」, 『日本統計学会誌』, 37 (2), 261-278.
6. 照井伸彦 (2024), “日本統計学会会長就任講演: ビジネスデータの統計モデリングとデータサイエンス”, 日本統計学会誌, 53, 247-273.
7. Akaike, H. (1980), ”A New Look at the Bayesian Approach,” *Biometrika*, 67(1), 117–123.
8. Allenby, G. M. and Rossi, P. E.(1996), ”Marketing models of consumer heterogeneity,” *Journal of Econometrics*, 89(1-2), 57–78.
9. Bass, F. M. (1969). ”A new product growth for model consumer durables,” *Management Science*, 15(5), 215–227.
10. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis (3rd ed.)*. CRC Press.
11. Rossi, P., G. Allenby and R. MaCulloch (2005), *Bayesian Statistics and Marketing*, Wiley.
12. Terui, N. and Wirawan Dony Dahana (2006a), “Estimating Heterogeneous Price Thresholds,” *Marketing Science*, 25 (4), 384-391.
13. Terui, N. and Wirawan Dony Dahana (2006b), “Price Customization Using Price Thresholds Estimated from Scanner Panel Data,” *Journal of Interactive Marketing*, 20(3), 58-70

「連続講義シリーズ：社会の中の統計科学」

欧州に波及する「レジスター型国勢調査」
～　統計手法・必要条件・課題　～

著者名　　千野　雅人（チノ　マサト）
所属・肩書　統計数理研究所 大学統計教員
　　育成センター長・特任教授

【はじめに】

2025 年は、5 年に一度の国勢調査の年である。

国勢調査は、国家の運営にも国民の生活にも、さらにいえば民主主義にも不可欠な統計調査である。各地域から選出される国会議員の数は、国勢調査の人口に従って定められる。実際、2020 年国勢調査の人口が 2021 年に公表された後、2022 年の公職選挙法の改正により、衆議院小選挙区の区割りが改定された。その結果、2024 年の衆議院議員選挙における小選挙区数は、5 都県で 10 増加する一方、10 県で 10 減少することとなった。国勢調査の結果が歪むと、近代的国家運営の基本原理とも言える民主主義が歪んでしまうのである。

国勢調査は、これまで多くの国で、調査員による全数実地調査の方法によって行われてきた。一方、全く新しい統計作成の方法が、欧州を中心に台頭してきている。それが、「レジスター型国勢調査」である。

【2020 年ラウンドから得た教訓】

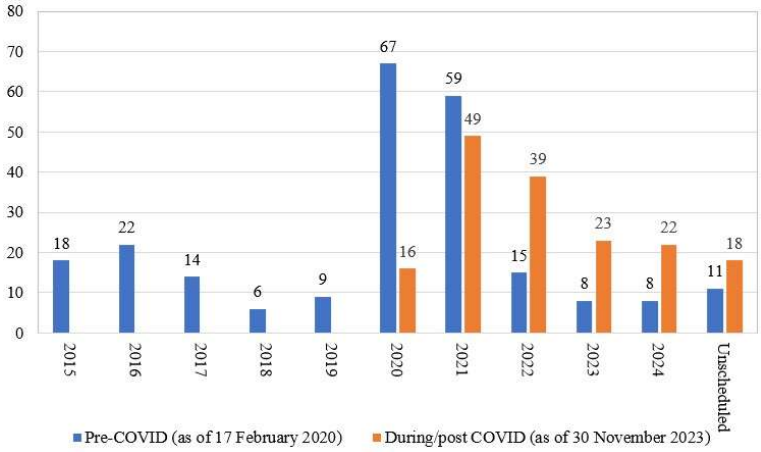
国連は、10 年ごとに、加盟各国に国勢調査の実施を促す計画を策定し、各国を支援している。2020 年ラウンド国勢調査計画では、2020 年を中心に、2015 年から 2024 年までの間に少なくとも 1 回の国勢調査を実施するよう加盟各国に勧告し、多くの国は、これに従って国勢調査を実施する計画であった。

しかし、2020 年ラウンドは、Covid19 パンデミックの到来期間と重なったため、多くの国が、調査実施年の延期、調査票回収期間の延長、調査員と世帯との直接接触を避ける調査方法への変更など、国勢調査の品質に負の影響をもたらしかねない措置を講じざるを得なくなった。

国勢調査の実施年をみると、パンデミック前の計画では、勧告に従って 2020 年に実施する国が最も多かった。しかし、パンデミック到来後には、多くの国が 2021 年以降に実施を延期した結果、2020 年実施国は少数派となり、2021 年に実施する国が最も多くなった（図１）。

パンデミックの状況は、国勢調査に多くの困難をもたらす一方で、統計作成プロセスにイノベーションの機会を生み出した。これを契機に、オンライン調査などインターネットを活用したデジタル・ファーストの調査手法の拡大や、統計作成の各段階での地理空間情報システムの活用、行政データの活用などの取組が、格段に進むことになったのである。このような中で特に注目を集めた動きの一つが、レジスター型国勢調査に関する取組である。

図１　2020 年ラウンド国勢調査を実施した
　　国の数（計画を含む）



出典：　United Nations　参考文献 [1]

【国勢調査の基本的な方法】

国勢調査の基本的な方法は、大きく次の３つに分けられる。

①最も多くの国が採用する「伝統的国勢調査」

②採用国が増加する「レジスター型国勢調査」

③これらを組み合わせた「複合型国勢調査」

世界の現状は、どうであろうか。

2019 年に国連が実施した加盟国調査の結果によれば、調査に回答した世界 158 か国のうち、①の方法を採用する国が 71%、②が 10%、③が 18%であり、現状では、①の「伝統的国勢調査」の方法を採用する国が、依然として大多数であることがわかる。

しかし、2018 年に国連欧州経済委員会が実施した欧州地域の加盟国調査の結果によれば、②の「レジスター型国勢調査」を採用する予定の国は、2020 年ラウンドの 13 か国から 2030 年ラウンドには 23 か国（調査に回答した欧州諸国の過半数）に増加するとされている。欧州を中心に、統計作成のパラダイムシフトが生じているようである。

以下、それぞれの方法を解説する。

【その 1 伝統的国勢調査】

伝統的国勢調査は、特定の基準日に、国内のすべての個人と世帯から、直接、必要な情報を収集する全数実地調査の方法であり、近代的国勢調査の開始当初から世界各国で行われてきた、国民にも国家統計局にもなじみ深い方法である。

調査方法には、①調査員が世帯を訪問し面接して聞き取った情報を調査票に記入する「他計式調査」と、②世帯員が自ら調査票に記入する「自計式調査」の 2 種類がある。日本の国勢調査は②の

自計式の方法で実施してきたため意外に感じるが、世界の多くの国の国勢調査は、①の他計式の方法により実施されている。

また、主要な設問のみから成るショートフォーム調査票と、必要なすべての設問から成るロングフォーム調査票の 2 種類の調査票を用意し、抽出された一部の世帯にのみロングフォーム調査票を配布する方法もある。

伝統的国勢調査の主な長所は、①各個人の情報を収集する「個別調査」であること、②国内すべての人を対象とする「普遍性」があること、③同じ時点で一斉に調査する「同時性」があること、④定期的な間隔で調査する「周期性」があること、⑤調査結果として「小地域統計」を得られること、という国勢調査の 5 つの本質的特徴をすべて満たしていることである。また、調査票の設問を変更することにより収集情報を機動的に設定できること、世界各国の統計局が豊富な経験を有する方法であること、なども利点といえる。

一方、最大の短所は、膨大なスタッフの確保と訓練を始めとする巨大事業の管理運営の複雑さと、コストの高さである。また、調査への回答率が継続して低下していること、タイムリーな調査結果を得がたいこと、十分な人数の調査員の確保が難しくなってきていること、なども難しい問題である。

【その 2 レジスター型国勢調査】

行政を適切に運営するために、行政機関が行政の規則に従って定義し収集した個人や住居に関する様々な情報（行政情報）を保管する記録簿・データベースを、行政レジスターという。行政レジスターの細部や記録プロセスは各国の国情によって異なるが、その大まかな種類は類似していると言われている。

レジスター型国勢調査は、個人、世帯、建物、税務、雇用など様々な分野の行政レジスターの情報を個人レベルで結合して構築した統計レジスターから、国勢調査に相当する統計を作成する方法であり、全数調査や標本調査のような実地調査の要素は全くない。

レジスター型国勢調査を実現するためには、主に次のような条件が満たされる必要がある。

・統計レジスターの構築に必要となる各種データの使用権限が、法令に明記されていること
・統計レジスターが、国内のすべての人口を網羅し、情報を継続的に更新する質の高いものであること
・各種行政レジスターの個別データの結合を可能とする個人識別番号の仕組みが存在すること
・各種行政レジスターのデータ品質の検証が可能であること（記録日時情報の存在など）

レジスター型国勢調査の主な長所は、長期的にみてコストがとても低いこと、高い頻度で国勢調査統計を提供できること、統計調査への回答負担がないこと、などが挙げられる。実地調査が全くないことにより、このような利点が得られるが、実は、レジスターを運用するシステムを構築するための初期コストは、決して低いものではない。

一方、最大の短所は、行政の目的に沿って収集された既存の行政レジスターの情報のみに頼らざるを得ないという事実である。データの定義や概念が国際統計基準と異なれば、国際比較可能性が

損なわれることになる。また、国勢調査統計の作成に必要となる一部の情報が、どの行政レジスターからも得られない可能性もある。

【その 3 複合型国勢調査】

複合型国勢調査は、行政レジスターの情報と、標本調査や全数調査などの実地調査から得られる情報とを組み合わせて、国勢調査の統計を作成する方法である。これにより、統計作成に必要となる情報の欠落のような行政レジスターの短所を補うことや、行政レジスターの情報の品質を検証し改善することなどができる。この方法は、世界の多くの国の支持を得るものになると予想されている。

複合の種類には、①全数実地調査と行政レジスター、②複合型国勢調査のために特別に実施する標本調査と行政レジスター、③既存の標本調査と行政レジスター、の 3 種類がある。

①の方法では、行政レジスターの個別の情報を各々の調査票にあらかじめ印刷することもできるが、その場合、十分なセキュリティを確保するための措置が必要となる。②の方法では、行政レジスターに欠落する情報を調査項目に設定するような柔軟な対応が可能となる。③の場合には、統計上必要な情報について相互に補完関係にある各種の標本調査と行政レジスターが既に存在することが必要となる。

複合型国勢調査の長所・短所は、①～③の方法ごとに異なるが、コスト、統計提供頻度、回答負担などにおいて、伝統的国勢調査とレジスター型国勢調査の中間にあると言える。

【行政レジスターの種類】

国勢調査統計の作成に用いる各種行政レジスターは、①基本レジスター、②専門分野レジスター、の２種類に区分することができる。

①の基本レジスターは、国勢調査の対象となる人や住居などの人口単位の全体を網羅し、そのストックに関する基本的な情報を保持するものである。それぞれの単位は、個人識別番号など固有の識別番号によって特定される。

典型的な基本レジスターは、人口レジスター、住宅レジスター、事業所レジスター、住所レジスターである。この中で特に重要な人口レジスターには、外国人を含め国内に普段居住するすべての人が、漏れなく重複なく記録されていることが理想である。しかし、不法移民、ホームレス、遊牧民など、登録に消極的だったり困難を伴ったりする人の記録は、漏れていることがある。逆に、登録を解除せずに帰国する外国人が誤って記録され続けたり、複数の住宅に居住して複数のレジスターに別人と識別された者の記録が重複したりすることがある。このような事態は、現実には避けられないことであり、注意が必要である。

②の専門分野レジスターは、特定分野の行政を執行するために人や住居に関する情報を保持するもので、このレジスターから、雇用、教育、収入など国勢調査統計の作成に必要となる情報を得ることができる。レジスターに記録される人や住居は、通常、基本レジスターと同一の固有識別番号によって特定され、各種レジスターの情報が結合される。また、必ずしも、国勢調査の対象となる

人口単位の全体を網羅するものではない。

典型的な専門分野レジスターは、健康保険、年金、税務、雇用、学生、社会保障施設入居、国境管理などの行政分野のレジスターである。

なお、国勢調査統計に必要となるすべての項目の情報を行政レジスターのみから収集することは、現実には不可能である。このため、労働力調査のような既存の標本調査の情報や、ビッグデータのような商業ベースの情報など、様々なデータ源の情報を活用することも必要となる。ビッグデータの場合には、データの偏り、他の情報と結合するための固有識別番号の欠如、コストの高さ、メタデータの欠如、特殊なファイル概念、などの問題に注意が必要である。

【レジスター型国勢調査の計画】

伝統的国勢調査からレジスター型国勢調査に移行しようとする場合、まずは複合型国勢調査に移行するなど、いくつかの段階を経て徐々に進めることが重要であり、通常は、数十年の期間を要する。しかし、使用できる行政レジスターの種類や品質が、統計作成の観点からは十分とは言えない可能性がある。このような場合には、何らかの形で伝統的国勢調査を実施し続けるべきであることを、当初から明確にしておくことも重要である。

レジスター型国勢調査に移行する際には、主に次の４つの段階を経ることになる。

① 必要な条件の確立・維持

② 行政レジスターなどのデータ源の特定

③ 各種データの統計レジスターへの変換

④ 統計の品質の管理と評価

①は、必要となる法令や制度、統計組織の能力などで、詳しくは後述する。

②では、既存の行政レジスター・データの概念・定義などを検証し、使用データを特定する。多くの行政データの概念・定義は、国勢調査に標準的に適用されるものとは異なることを認識する必要がある。そのような相違が許容されるか否かは、実地調査経費の削減と統計の連続性確保とのバランス、利用者のニーズ、適時性など、様々な要因を考慮して判断する。また、行政レジスターのデータのみでは不十分な場合には、既存の標本調査や特別実施の統計調査など、他のデータ源の活用も検討する。

③では、固有識別番号などによって各種データ源の個別データを結合し、国勢調査統計を作成するための人口統計レジスターを構築する。

行政レジスターの人口データは、行政目的に従って法令規則を基に登録されるが、人口統計レジスターには、統計の定義・概念を基に国内居住者のデータのみが登録されるようにする。行政への登録を解除せずに国内外に移動した者や死亡者など、既に居住者でなくなった者のデータが含まれないようにするため、年金支給レジスターを始めとする複数のレジスター情報を確認するなど、いくつかの「活動ルール」を設けて登録データを検証する（「生命兆候法」”Signs-of-Life method”）。

このほか、複数の行政レジスターに別人として登録されてしまった者や、更新時期の相違による新旧情報の混在など、レジスター間の矛盾を解消する。また、誤ったデータを訂正するための審査

や、欠損データに統計上整合的なデータを挿入する補定などの統計編成処理を行う。

④は、データ源、入力データ、データ処理プロセス、統計成果物、の各段階で品質の管理と評価を行うものである。新たなデータ源を活用するためには、テストデータの取得・分析などにより、事前にデータの品質を評価することも重要である。

【レジスター型国勢調査の必要条件】

レジスター型国勢調査に移行するために必要となる条件は、次の 4 つである。

①長期戦略ビジョンの確立

②政治レベルの支持と法的枠組み

③国民の支持と利用者の関与

④その他運用上の条件

①は、レジスター型国勢調査への移行が国家統計局の統計作成プロセス変革の第一歩に過ぎず、長期的には、より広範なレジスター型統計作成システムへの移行を目指す、という長期のビジョンを確立することである。レジスター型国勢調査の統計作成システムを開発し導入するための初期費用や資源は、レジスター型統計作成システムがより広範な統計活動に活用されることによって、正当化される。

②は、レジスター型統計作成システムに移行するために必要となる法改正や、国家統計局へのデータ提供を困難にする障壁の除去などに、政治レベルの支持が欠かせない、ということである。また、国家統計局が行政データを利用する権限、様々な情報源の個別データを結合する権限、行政データの作成・訂正・削除に何らか関与する権限などが、法令に明記されることが必要である。逆に、国家統計局は、行政データを統計目的にのみ利用し、その秘密を保護することなどについて、法的義務を負うべきである。さらに、行政の個別データを統計目的で利用することがあっても、統計目的で収集した個別データを行政目的に利用してはならないという、「一方通行原則」（"one-way traffic only principle"）を遵守することも必要である。

③は、個別データの結合などのレジスター型統計の作成方法について、必要となる法令上の権限を有するだけでなく、多くの国民の支持を得ることが極めて重要ということである。このためには、個別データが統計目的のみに利用され、その秘密が厳格に保護されることについて、国民の十分な理解を得ることが必要である。また、コストの削減、適時性の向上、回答負担の軽減など、統計作成方法の変更による利点が、時系列の途絶や統計の欠落などの欠点を上回るかどうかについて、利用者を始めとする関係者の意見を聴いて判断することが重要である。

④は、行政レジスター所管機関と国家統計局との間の良好な協力関係の構築、行政レジスターの情報に関する十分な知識の獲得、各種レジスターの個別データに共通する固有識別番号の存在、国家統計局職員の専門的技術能力の開発などが必要であるということである。

【レジスター型国勢調査の課題】

レジスター型国勢調査の最大の課題は、統計の作成が行政レジスターの所管機関に大きく依存す

ることになる点である。所管行政機関にとって統計作成は、優先される中核的な活動ではないため、行政データの提供に遅れが生じたり、政策の変更によってデータの概念・定義が変更されたりする恐れがある。また、失業者を始めとする人口データの概念・定義は、国勢調査に国際的に適用されるものとは異なることが多い。

さらに、行政レジスターごとにデータの概念・定義や登録の時期が異なること、登録や入力に誤りがあることなどにより、行政レジスター間でデータの不整合が生じることがある。また、共通する固有識別番号の仕組みがない場合には、統計の作成に必要となる各種データ源の個別データの結合は、極めて難しくなる。

【まとめ】

レジスター型や複合型の国勢調査は、より頻繁でタイムリーな統計提供、コストや回答負担の大幅な削減などの利点があり、特にパンデミック後の世界の潮流となっている。一方で、国家統計局が行政データの概念・定義などに関して所管外であることや、個別データの統計利用のための法令枠組が必要であることなど、解決すべき課題も多い。

レジスター型国勢調査への移行には、各国共通の理想的な方法や期間は存在しない。公的統計作成の様々な段階で行政データを活用する広範な統計作成システムの構築に向け、法令枠組や固有識別番号の整備状況などの国情に応じて、段階的に取組を進めることが肝要である。

（この原稿は、「ESTRELA 2025 年 7 月号」（Sinfonica）への寄稿原稿を加工したものである。）

＊参考文献

[1] “Population and housing censuses : Report of the Secretary-General E/CN.3/2024/15” United Nations Economic and Social Council, Statistical Commission Fifty-fifth session
https://unstats.un.org/UNSDWebsite/statcom/session_55/documents/2024-15-Censuses-E.pdf

[2] “Handbook on Registers-based Population and Housing Censuses, Version: December 2022” United Nations Department of Economic and Social Affairs, Statistics Division
https://unstats.un.org/unsd/demographic-social/publication/handbook-registers-phc.pdf

[3] “Principles and Recommendations on Population and Housing Censuses, Revision 3, 2017” United Nations Department of Economic and Social Affairs, Statistics Division
https://unstats.un.org/unsd/publication/SeriesM/Series_M67Rev3en.pdf

生命保険料決定と公正なデータサイエンス

吉田　靖 *

2025 年 10 月

概要

要約

本論文は，生命保険料の決定過程における数理的合理性と法的・倫理的正当性の両立を中心課題としている．AI や機械学習の導入によって，死亡率推計とリスク細分化はかつてない精度で実現可能となったが，その一方で「統計的差別」と「不当な差別的取扱い」との境界が曖昧になりつつある．本稿は，生命保険の数理モデルを単なる技術的問題として扱うのではなく，社会的制度としての正当性の観点から再評価し，予測の合理性と規範的正当性の統合を目指し，公正なデータサイエンスの拡充を目的としている．

1　はじめに

近年，人工知能（AI）や機械学習の社会的利用は急速に普及しつつある．保険業も，インシュアテックということばに表されているようにその例外ではなく，Bauer et al. [4] は，機械学習による保険の不正検知など広く学術論文を紹介している．保険業の中でも，特に生命保険料の決定に関しては古くから数理的手法が用いられ，現在ではニューラルネットワークや機械学習によるモデルも多く提案されている．また，契約後の健康診断情報やスマートフォンなどのセンサー情報に基づいて，保険料が動的に変化する保険商品も登場し，医療においてはヒトゲノム解析の進展により遺伝子情報の利用が広まっている．

その一方で，欧州では従来は合理的な差別として認められていた保険契約における性差別

* 東京経済大学教授，統計数理研究所客員教授

を禁止する EU Test-Achats 判決[*1] が下されるなど，データサイエンス技術の開発・利用において法的・倫理的な公平性の配慮がより重要になっている．このように，機械学習モデルによる料率決定がいかに統計的に高精度であっても，その基礎となる変数選択や予測構造が法的・倫理的に公平であるとは限らないという問題が存在する．機械学習モデルが予測する死亡率の「差異」が，社会的に許容される「差別」か，それとも「不当な差別的取扱い」に該当するのかは，法的には憲法から保険法・保険業法・個人情報保護法までが交錯する論点である．本稿は保険契約者の属性に基づく保険料算定の基礎である死亡率の推計の問題について，推計モデルの学術的な現状を紹介すると共に，法的・倫理的な面の課題を検討する．

本稿は，保険数理と法規範を単に並列するのではなく，数理モデルの正当性を法的規範によって再評価するものである．すなわち，数理的に正しいことが社会的に正義であるとは限らないため，保険という社会制度の下では，予測の合理性と規範的正当性の両立が求められる．データサイエンス技術の進展によって可視化された差異を，どこまで「公正な差別」として許容しうるか——それが公正なデータサイエンスの拡充を目指す本稿の中心的な問いである．

本稿の構成は次のとおりである．つづく第 2 章では，今後の議論のために，二大原則といわれる収支相等の原則および給付反対給付均等の原則の制度的・アクチュアリアルな意義と平準純保険料方式の数理構造を再確認する．第 3 章では，Lee – Carter モデル以降の死亡率予測研究の発展を概観し，ニューラルネットワークを用いた近年の手法である Richman and Wüthrich [12] などを紹介する．第 4 章では，Arrow [1]，Mossin [9]，Yaari [14] および Rothschild and Stiglitz [13] らを基礎とする生命保険の経済学的理論を整理し，効率性と公平性のトレードオフ構造を示す．第 5 章では，EU Test-Achats 判決による性別による保険料差別禁止に注目し，「統計的に正しいこと」と「法的・倫理的に正当であること」が必ずしも一致しないことを議論する．最後に第 6 章で，生命保険料決定の課題として，説明可能性（explainability）とアルゴリズム的公平性（algorithmic fairness）を取り上げ，Barocas and Selbst [3]，Barocas et al. [2] および Charpentier [5] の議論を参照しつつ，公平なデー

[*1] European Court of Justice [6]

タサイエンスを検討する．

2　保険料決定の基礎：収支相等の原則と給付反対給付均等の原則と保険数理

保険料の決定に関し，一般に保険の二大原則と呼ばれる収支相等の原則と給付反対給付均等の原則があるとされている[*2]．「収支相等の原則」の定義は公益社団法人日本アクチュアリー会（以下，アクチュアリー会という）のウェブサイト「アクチュアリーの用語集」[*3]によれば「保険の種類ごとや保険を引き受けている集団ごとに「保険料・掛金・運用益等による収入総額」が「保険金・年金・経費等による支出総額」と一致するように，保険料・掛金を算定する原則」と示されている．「給付反対給付均等の原則」については「アクチュアリーの用語集」に記載はないが，一般社団法人生命保険協会（以下，生命保険協会）ウェブサイト「生命保険の基礎知識」[*4]では，「リスクの高さに応じて保険料を算出することで保険契約者の負担は公平となります．この原則を「給付反対給付均等の原則」といいます.」と説明されている．「均等」という用語が示すように，これは，個々の保険契約において，保険料（反対給付）と保険金（給付）の経済的価値が社会的に許容可能な範囲で公平となるべきという原則である．法令上は，保険業法[*5]第 3 条第 1 項により，「保険業は，内閣総理大臣の免許を受けた者でなければ，行うことができ」ず，第 4 条第 2 項第 4 号に免許申請書には「保険料及び責任準備金の算出方法書」の添付が必要とされている．その算出方法書に関しては，第 5 条第 1 項第 4 号に基準として「イ　保険料及び責任準備金の算出方法が，保険数理に基づき，合理的かつ妥当なものであること．ロ　保険料に関し，特定の者に対して不当な差別的取扱いをするものでないこと.」と規定されている．さらに，第 116 条第 2 項では「長期

[*2] 日本の法令では「収支相等」および「給付反対給付均等」という語は用いられていない．保険学的には，二大原則の法的位置づけに関しては議論があるが，詳細は堀井 [20]，舩津 [19] などに記載がある．

[*3] https://www.actuaries.jp/actuary/glossary.html

[*4] https://www.seiho.or.jp/data/billboard/introduction/content03/

[*5] 保険業法は，「保険業の公共性にかんがみ，保険業を行う者の業務の健全かつ適切な運営及び保険募集の公正を確保することにより，保険契約者等の保護を図り，もって国民生活の安定及び国民経済の健全な発展に資することを目的」と規定されている．

の保険契約で内閣府令で定めるものに係る責任準備金の積立方式及び予定死亡率その他の責任準備金の計算の基礎となるべき係数の水準については，内閣総理大臣が必要な定めをすることができる」と規定され，これに基づき，大蔵省告示第 48 号 (平成 8 年 2 月 29 日) では，責任準備金の積立方式は平準純保険料式とすることや，予定死亡率はアクチュアリー会が作成した生保標準生命表[*6]の死亡率と定められている．

また，保険契約者に対しては保険法第 37 条で「保険契約者又は被保険者になる者は，生命保険契約の締結に際し，保険事故（被保険者の死亡又は一定の時点における生存をいう．以下この章において同じ．）の発生の可能性（以下この章において「危険」という．）に関する重要な事項のうち保険者になる者が告知を求めたもの（第 55 条第 1 項及び第 56 条第 1 項において「告知事項」という．）について，事実の告知をしなければならない．」と規定され，この告知義務制度は，情報の非対称性を是正するものとして機能していると考えることができる．この告知事項に関して，金融庁「保険会社向けの総合的な監督指針（IV ．保険商品審査上の留意点等）IV―１―５　告知項目」[*7]では「保険契約者又は被保険者に求める告知項目は，保険会社が危険選択を行う上で必要なものに限定されているか．また，「趣味」など判断基準があいまいな用語は適当でないことに留意するものとする．」とのみされていて，具体的な制限を課していない．

ここで，前述の大蔵省告示第 48 号 (平成 8 年 2 月 29 日) により，責任準備金の積立方式は平準純保険料式とすることと，予定死亡率はアクチュアリー会が作成した生保標準生命表の死亡率と定められていることによる保険料の決定方式について確認しておく．

まず，生命保険における「平準」とは，時間の経過とともに変化するリスクや支払額を平均化し，契約期間を通じて保険料一定化させることを指す．通常，生命保険では被保険者の死亡リスク（死亡率）は年齢とともに上昇する．したがって，そのままリスクを保険料に反映すれば，年ごとに保険料が増加することとなり，高齢になるほど保険料が高額になり，保険料の支払いが困難になることも考えられる．これを避けるため，契約期間を通じて一定額の保険料を支払う方式が採用されている．この場合，契約初期に支払われる保険料のうち，

[*6] https://www.actuaries.jp/lib/standard-life-table/index2018.html
[*7] https://www.fsa.go.jp/common/law/guide/ins/04.html#04_01

その期間のリスクに対して過剰な部分を責任準備金として積み立て，契約後期のリスク増大に伴う不足分を責任準備金から取り崩して補う仕組みとなる．

単純な定期保険[*8]を例に，平準保険料の数理構造を次に示す．保険数理[*9]上，採用した生命表の示す死亡率を予定死亡率，利息の計算に使用する利率を予定利率，保険制度の運営に必要な経費の保険金額あるいは保険料に対する比率を予定事業費率, という．これら 3 つの率が将来変動することに備えて，予定死亡率はやや高め，予定利率はやや低め，予定事業費率はやや高めに設定され，予定死亡率と予定利率から計算される保険料を純保険料，予定事業費率から計算される保険料を付加保険料という．純保険料と付加保険料を合計したものが契約者が払い込む営業保険料である．

いま，被保険者の年齢を x，保険期間を n 年，保険金額を S，利率を i とすると，各年の純保険料は死亡率 q_{x+t-1} に比例して変化し，これを契約期間全体で平準化した一定の純保険料 P は次式で表される．

$$P = \frac{S\sum_{t=1}^{n} v^t \, q_{x+t-1} \, {}_{t-1}p_x}{\sum_{t=1}^{n} v^{t-1} \, {}_{t-1}p_x}. \tag{1}$$

ここで，$v = 1/(1+i)$ は割引率，${}_{t-1}p_x$ は x 歳から $x+t-1$ 歳まで生存する確率である．

このように契約時の年齢により保険料は異なり，さらに「標準生命表 2018」は男女で異なるものになっている．ただし，日本アクチュアリー会 [17] にあるように「標準生命表は，健全性確保の観点から保険業法において積立が求められる責任準備金の計算の基礎であり，保険料計算の基礎である予定死亡率とは性格が異なりますので，価格については各保険会社の経営判断によって決定され」ている．また，作成に際して，各種の補正を行っているが，基本的には生命保険会社 29 社の 2008，2009，2011 観察年度[*10]の実績を基礎データとし，国民生命表により 2018 年までの時点修正を行っているが，有診査で年齢別・男女別のみのデータに基づいている．

実際の生命保険料 (営業保険料) についてはウェブサイトで，性別と生年月日を入力して

[*8] 一定の保険期間を契約時に定め，その期間に死亡した場合に保険金を受け取れる生命保険

[*9] たとえば二見 [18]

[*10] 東日本大震災の影響を除くため 2010 観察年度は除かれている．

見積もることが可能な例が多いが，純保険料と付加保険料を公表している例は少なく，ライフネット生命のウェブサイト[*11]によれば，2024 年 10 月現在で, 同社の定期死亡保険「かぞくへの保険」（保険期間 10 年）で保険金額 1,000 万円，40 歳の場合，男性は純保険料 1,365 円（月額，以下同じ），付加保険料 560 円，女性は純保険料 1,463 円，付加保険料 474 円, 50 歳の場合，男性は純保険料 3,233 円，付加保険料 984 円，女性は純保険料 1,985 円，付加保険料 701 円などとなっている．このように性別や年齢により保険料は大きく異なる．

3　死亡率予測モデルの進展

Lee and Carter [7] による Lee–Carter モデルは，死亡率の時間的変化を単純かつ柔軟に捉えるために考案された代表的な人口統計モデルであり，現在では様々な改良が加えられて生命保険数理，長寿ボンドのプライシングや人口予測の分野で広く利用されている．モデルの基本的アイデアは，年齢 x と暦年 t に依存する対数死亡率 $m_{x,t}$ を，「年齢効果」，「時系列要因」，「感応度」の 3 要素に分解して表現することである．

Lee–Carter モデルの中心的な式は式 (2) のとおりである．

$$\ln m_{x,t} = a_x + b_x k_t + \varepsilon_{x,t}. \tag{2}$$

ここで各項の意味は以下のとおりである．

$m_{x,t}$：年齢 x における暦年 t の死亡率.

a_x：年齢 x に固有の平均的な対数死亡率（年齢効果）.

b_x：年齢 x における死亡率の変化に対する感応度.

k_t：時点 t における全体的な死亡水準の変動を示す時系列要因.

$\varepsilon_{x,t}$：残差項（独立同分布を仮定）.

このモデルは，対数死亡率を時系列要因 k_t と各年齢の感応度 b_x の積で表現することによ

[*11] https://www.lifenet-seimei.co.jp/shared/pdf/insurance_table_202410.pdf

り，死亡率の年齢構造を保持しつつ全体的な水準変化を表現する構造となっている．

式 (2) において，a_x, b_x, k_t は未知パラメータである．通常，以下の手順で推定が行われる．

まず，観測された対数死亡率 $m_{x,t}$ の平均を年齢方向にとり，a_x とする．

$$a_x = \frac{1}{T}\sum_{t=1}^{T} \ln m_{x,t}. \tag{3}$$

次に，a_x を除いた残差部分

$$\ln m_{x,t} - a_x \tag{4}$$

に対して特異値分解（SVD）を行い，第 1 主成分を b_x，k_t に対応させる．

さらにモデルの同定（識別）を行うために，式 (5) の制約条件を課す．

$$\sum_x b_x = 1, \qquad \sum_t k_t = 0. \tag{5}$$

これにより，死亡率行列の主要な変動を 1 次元の時系列 k_t と感応度 b_x によって説明することができる．

推定された k_t は時系列的に滑らかなトレンドを示すため，一般に ARIMA モデルなどで外挿される．将来の死亡率は，推定された a_x, b_x と時系列モデルにより外挿された $\hat{k}_t$ を用いて式 (6) で予測される．

$$\ln \hat{m}_{x,t} = a_x + b_x \hat{k}_t. \tag{6}$$

このように Lee–Carter モデルの特徴は，構造が比較的単純でパラメータ数が少なく，死亡率全体の動向を少数の要因で把握できる点にある．

その後，Lee–Carter モデルに対して様々な改良が提案されいる．たとえば，Renshaw and Haberman [11] は Lee–Carter モデルにコホート効果（cohort effect）を導入し，死亡率の世代間の違いを説明するための枠組みを提案している．また，Li and Lee [8] は複数集団の死亡率を同時にモデル化する multi-population mortality model を提案し，死亡率トレン

ドの共通性を活用することで，個別地域の死亡率予測の精度を高めることができることを示している．

近年では，機械学習（machine learning）および深層学習（deep learning）技術を死亡率モデリングに応用する研究が急速に進展し，Richman and Wüthrich [12] はニューラルネットワークを用いて Lee–Carter モデルを複数集団へ拡張し，最適なモデル構造を自動選択する方法を提案している．彼らはこのモデルを Human Mortality Database (HMD) のすべての国における 1950 年以降の死亡率に適合させ，予測性能に極めて競争力があることを確認している．さらに，Perla et al. [10] は比較的単純な浅層畳み込みネットワークモデルを用いて Lee – Carter モデルを一般化し，深層ネットワークモデルは浅層畳み込みネットワークモデルよりも予測性能の向上にはつながらないことを示している．

このように，ニューラルネットワークを用いた死亡率モデリングは，従来の線形モデルの枠組みを自然に拡張するものである．

4　生命保険の経済学的基礎

個人が生命保険を契約する理由は，経済学的には期待効用理論 に基づいて説明される．生命保険契約は，死亡という不確実事象に対するリスク回避的な資源配分 の一形態であり，死亡時と生存時の富の分布を調整することで，個人または家族全体の効用を最大化する行動として理解される．

以下で，Arrow [1] と Mossin [9] による期待効用理論と保険契約，Yaari [14] による異主体効用モデルによる生命保険の特性，および Rothschild and Stiglitz [13] による逆選択を整理する．

まず，個人は，生存状態 L と死亡状態 D の 2 つの状態における富 W に対して効用関数 $U(W)$ を持ち，効用関数として，次のように限界効用逓減（risk aversion）を仮定する．

$$U'(W) > 0, \quad U''(W) < 0. \tag{7}$$

個人が保険料 (営業保険料) p を支払い，死亡時に保険金 B を受取人が受け取る生命保険

契約を考える．この個人すなわち被保険者の死亡確率を q，生存確率を $(1-q)$ とすると，生命保険契約時の期待効用は次式で表される．なお，ここでは簡単化のため，貨幣の時間価値は無視できるものとする．

$$E[U] = (1-q)\,U(W-p) + q\,U(W-p+B). \tag{8}$$

保険に加入しない場合の期待効用は，

$$E_0[U] = (1-q)\,U(W) + q\,U(W-L). \tag{9}$$

である．ただし L は死亡により生じる経済的損失（受取人の所得減少など）を表す．

この個人が生命保険を契約する条件は次の不等式で与えられる．

$$E[U] > E_0[U]. \tag{10}$$

リスク回避的な個人は，富の変動よりも安定性を好むため，保険料 p が純保険料より高くても保険契約を選好する．純保険料は，期待値に基づき次のように表される．

$$p^* = qB. \tag{11}$$

実際の保険料には，付加保険料（loading）を含むため，$p > p^*$ であっても，保険契約時の期待効用が契約前の期待効用を上回る場合には，個人は保険契約を選択する．

さらに生命保険の特徴は，保険金の受取主体が本人ではない点にある．このため，本人の効用最大化問題は，受取人の効用を自身の効用関数に内生化する形で表現される．

Yaari [14] は，生存時消費 C_L と死亡時消費（受取人消費）C_D に基づく期待効用を次のように定式化している．

$$E[U] = (1-q)U(C_L) + qV(C_D). \tag{12}$$

ここで $U(\cdot)$ は本人の効用，$V(\cdot)$ は受取人の効用を表す．保険契約は C_D を増加させ，C_L

を減少させるが，利他的な個人は受取人の生活水準向上を自己効用の一部として評価する．

より現実的には，保険会社は個人のリスクタイプ（死亡確率 q）を完全には観察できないため，Rothschild and Stiglitz [13] が指摘するように情報の非対称性が存在する．この状況で保険料が平均的リスクを基準として設定される場合，低リスク者が市場から退出し，高リスク者のみが残る逆選択（adverse selection）が発生する．これを防ぐために保険者はリスク区分を細分化し，個人属性に応じた料率を設定する．これは数理的には収支相等原則を維持する合理的行動であるが，「特定の者に対する不当な差別的取扱い」としての問題が発生する可能性を孕む．したがって，経済学的効率性を追求するほど，公平性との間に緊張関係が生じる構造が保険契約にはある[*12]．

この構造は，データサイエンスが導入された現代において，より複雑な形で再現される．AI モデルは，数百万件の過去データから潜在的なリスク要因を抽出し，個々の契約者に対して精密なリスクスコアを算出することができる．このとき，モデルの目的関数は統計的誤差（loss function）の最小化に基づくため，経済学的に言えば効率性の極大化を志向している．しかし，AI が予測する死亡率が社会的属性や遺伝情報と強く相関している場合，それは統計的には最適でも，法的・倫理的には不当な不均衡を生じうる．したがって，予測モデルによる効率性の向上が，必ずしも社会的公平性の実現を意味しない点に留意する必要がある．死亡率の予測問題は，単なる統計的推定ではなく，この社会的公平性をどう考慮するかという制度設計問題として理解されるべきである．すなわち，保険料決定の正当性は，技術的最適化の問題ではなく，「社会がどのような効率性・公平性の組合せを正当と認めるか」という規範選択の問題である．特に，遺伝情報を料率に反映させることは，「遺伝差別（genetic discrimination）」の問題を引き起こすおそれがあり，倫理的にも法的にも極めて慎重な取扱いが求められる．日本では生命保険協会 [15] にあるように，現在の取扱としては「遺伝学的検査結果の収集・利用は行って」いないものの，「医療の進歩や社会的な議論の成熟等，環境や情勢の変化に応じ，特に今後ゲノム医療が普及し遺伝情報について消費者の正確な理解が

[*12] ただし，大倉 [16] は「ハイリスクタイプ消費者とローリスクタイプ消費者との事故発生確率の差が小さいとき，あるいはハイリスクタイプ消費者の人数が相対的に少ないとき，規制等によってリスク細分化を禁止することが，パレートの意味において望ましいことを」経済学的に証明している．

進むことに伴い，新たな課題が認識された場合等には，監督官庁の指導と医療・医学等の関係者の意見を参考とし見直しを行うことを含め適時適切に対応して参ります.」としている．

さらに，データ倫理（data ethics）と個人情報保護の問題がある．前述のようにリスク細分化が進んだ死亡率予測モデルは多種多様な個人データを入力変数として用いるが，その中には，健康診断結果，ライフログ，遺伝情報など，個人情報保護法に定める個人情報や要配慮個人情報が含まれる可能性がある．これらを保険料決定に利用することは，個人情報保護法上の適法性・必要性・相当性の観点からも慎重に検討されるべきである．

5　EU における性別保険料の禁止判決

本節では公平性の議論に資するため，EU における性別保険料の禁止判決について紹介する．

2004 年に施行された EU 男女平等指令（Directive 2004/113/EC）は，財・サービスの供給における性別による差別を原則として禁止していたものの，生命保険や年金などの分野では，「統計的に性別がリスクに影響する」場合に限り，加盟国が性差による保険料区分を一時的に認める「例外条項」が設けられていた．しかし，ベルギーの消費者団体 Test-Achats が「性別による保険料差別は EU 基本権憲章に反する」として提訴したところ，European Court of Justice [6] にあるように欧州司法裁判所は，性別は保険料を決める正当な理由にはならないとして「統計的に差がある」ことを根拠に無期限の例外を認めるのは，基本権憲章に定められた男女平等原則に反すると判断した．これにより，2012 年 12 月 21 日以降に締結されるすべての新規保険契約について，性別に基づく保険料・給付の差を設けることは禁止された．この判決に対し，保険業界は「リスクに見合わない価格設定」になると批判しましたが，EU は人権と平等の原則を優先し，「個人を性別という集団の統計に基づいて扱うことは不当な差別」と位置づけ，「個人単位のリスク評価」が重視されるべきとしました．

6　おわりに：データサイエンスによる料率設計と倫理的課題

近年のデータサイエンスの発展により，生命保険料の決定過程は大きく変容している．従来の料率算出は，予定死亡率・予定利率・予定事業費率といった基礎率を用いたモデルに依拠していたが，今日では，機械学習モデルによって被保険者の健康情報，ウェアラブル端末のデータなど，多次元の変数を用いた動的なリスク推定が可能となっている．これらのモデルは，従来型の一般化線形モデル（GLM）を拡張し，非線形相互作用を自動的に学習する Gradient Boosting，Random Forest，Neural Network などの手法が用いられる．特にディープラーニングを活用した健康リスクスコアの算定は，生命保険料の個別化（personalized pricing）を現実のものとしている．

しかし，このような AI による料率設計は，法的・倫理的な観点から多くの課題を提起している．その第一として，説明可能性（explainability）の問題がある．機械学習モデルは高い予測精度を有する一方で，その意思決定過程がブラックボックス化しやすい．たとえば，ニューラルネットワークが算出したリスクスコアがどの要因に起因するかを人間が直感的に理解することは困難である．このため，契約者に対して「なぜこの保険料になったのか」を説明することが難しく，これは契約上の説明義務・適合性義務に関わる問題であり，消費者契約法および保険法上の「契約者保護」の理念に抵触し得る．

こうした課題に対処するため，説明可能 AI（XAI）やアルゴリズム的公平性を保険数理に導入する試みが進んでいる．代表的な XAI 手法である SHAP（Shapley Additive Explanations）や LIME（Local Interpretable Model-agnostic Explanations）は，各変数が予測結果に与える寄与度を可視化し，AI の意思決定過程を人間が理解可能な形に変換する．

さらに，こうした問題を理論的に分析するために注目されるのが，Barocas and Selbst [3] および Barocas et al. [2] による Algorithmic Fairness の議論である．彼らは，アルゴリズムによる意思決定が社会的に不当な格差を生じさせることを防ぐため，複数の数理的公平性指標（demographic parity, equal opportunity, calibration）を提示している．

まず，Barocas and Selbst [3] は，ビッグデータによる自動化が既存の社会的不平等を再生産する可能性を「Big Data's Disparate Impact」として指摘した．彼らは，差別が意図的でなくとも，モデルの特徴量やデータ収集過程に構造的バイアスが含まれる場合，結果的に属性集団間で不当な不利益を生じうると論じている．次に Barocas et al. [2] では，この問題を定式化するために次の 3 つの公平性概念を提示している．すなわち，(1)demographic parity：属性群間で予測結果の分布が等しいこと，(2)equal opportunity：真にポジティブな対象に対する真陽性率が群間で等しいこと，(3)calibration：予測確率と実際結果の対応関係が群間で等しいことの 3 点である．そしてこれらは相互に両立し得ないことを示し，どれを採用するか自体が規範的選択を伴うので，モデル設計者はどの公平性を優先するかを倫理的・社会的判断として明示すべきとしている．

一方，Charpentier [5] は AI や機械学習を用いた保険料算定が急速に普及する中で，「どのような差別が正当化され，どのような区別が不当なのか」という問題について包括的に整理した近年の代表的文献である．Charpentier [5] は，保険における *fair discrimination*（公正な差別）という逆説的概念を中心に，統計的・倫理的・法的・アクチュアリアルな次の 4 つの点を論じている．すなわち，（1）GLM や GAM からニューラルネットワークまでの保険数理における予測モデリング技術を概観したうえでの，保険料算定における「区別」と「差別」の境界，（2）個人情報保護法制（GDPR など）と保険データ利用の倫理的側面を分析し，欠落変数バイアスやシンプソンのパラドックス等の統計的偏りが公平性判断に及ぼす影響，（3）グループ公正（demographic parity, equalized odds, calibration）と個人公正（similarity, counterfactual fairness, optimal transport）を数理的に定義して比較，（4）不公正を是正するための前処理・学習中処理・後処理技術を整理し，GermanCredit, FrenchMotor などのデータで実証することである．したがって，保険数学，データサイエンス，倫理学，および法政策の各分野を橋渡しし，AI 時代のアクチュアリーが直面する「統計的差別と社会的公平性」のジレンマに理論的基盤を与えている．特に，保険数理の根幹であるリスク細分化をめぐる倫理的判断を *discrimination versus fairness* の軸で再構成しており，リスク細分化と差別禁止原則の両立可能性を検討するうえで，同書の議論は理論的基

盤を与えるものになり得る．

さらに，AI 料率の利用には「二次的不公平」のリスクがある．すなわち，初期学習データにバイアスが含まれている場合，AI がその偏りを強化し，新しいデータを再び不均衡な構造で学習してしまう「バイアスの自己増殖」が起こり得る．これを防ぐには，モデルの定期的な再訓練と公平性評価を制度的に義務づける仕組みが必要である．

このように，データサイエンスの導入は保険料決定の精度を高める一方で，法的・倫理的課題を増幅させるという二面性を持つ．保険業法が要求する「合理的かつ妥当な算出方法」（第 5 条第 1 項第 4 号イ）は，AI 時代においては単なる数理的合理性だけでなく，「説明可能で社会的に受容可能な合理性」へと拡張される必要がある．

以上のように本稿では，網羅的ではないが，重要と思われる文献を紹介してきた．生命保険会社が健康増進型保険と呼んでいる保険商品では健康状態による保険料の差異が自分の健康状態を管理するインセンティブになる可能性もあるがこの面での研究は少ない．特に実証的な研究による検証が望まれる．また，日本では生命表そのものは必ずしも広く一般には浸透していないかもしれないが，男女で平均寿命が異なることが長年にわたり日常で話題になっており，数値的にも明確になっている歴史がある．一方，ゲノム情報は一般的に寿命に与える計量的な影響はいまだ一般的になったとはいえず，その精度も明確でないといえよう．個人のゲノム情報も本人が望めば 3 万円程度の費用である程度の把握ができるようになったとはいえ，大半の人々はまだ利用していない．これらの問題が解消されることが，「説明可能で社会的に受容可能な合理性」の議論に寄与すると思われる．

謝辞

本稿は，統計数理研究所一般研究 1(2023- ISMCRP-1006) および東京経済大学 2024 年度個人研究助成費（研究番号：24-31）による研究成果の一部である．

参考文献

[1] Arrow, K. J. (1963). "Uncertainty and the welfare economics of medical care." *American Economic Review*, 53(5), 941 – 973.

[2] Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. https://fairmlbook.org/.

[3] Barocas, S., and Selbst, A. D. (2016). "Big data's disparate impact." *California Law Review*, 104(3), 671 – 732.

[4] Bauer, D., Schmit, J., and Sydnor, J. (2024). The economics of emerging insurance technologies: Theory and early evidence. *Journal of Risk and Insurance*, 91, 809–812. https://doi.org/10.1111/jori.12495.

[5] Charpentier, A. (2024). *Insurance, Biases, Discrimination and Fairness*. Springer Actuarial Series, Springer, Cham. https://doi.org/10.1007/978-3-031-49783-4.

[6] European Court of Justice (2011). Case C-236/09, *Association Belge des Consommateurs Test-Achats ASBL and Others v Conseil des ministres*, Judgment of 1 March 2011.

[7] Lee, R. D., and Carter, L. R. (1992). "Modeling and forecasting U.S. mortality." *Journal of the American Statistical Association*, 87(419), 659 – 671.

[8] Li, N., and Lee, R. D. (2005). "Coherent mortality forecasts for a group of populations: An extension of the Lee – Carter method." *Demography*, 42(3), 575 – 594.

[9] Mossin, J. (1968). "Aspects of rational insurance purchasing." *Journal of Political Economy*, 76(4), 553 – 568.

[10] Perla, F., Richman, R., Scognamiglio, S., and Wüthrich, M. V. (2021). Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal*, 2021(7), 572 – 598. https://doi.org/10.1080/03461238.2020.1867232.

[11] Renshaw, A. E., and Haberman, S. (2006). “A cohort-based extension to the Lee – Carter model for mortality reduction factors.” *Insurance: Mathematics and Economics*, 38(3), 556 – 570.

[12] Richman, R., and Wüthrich, M. V. (2021). “A neural network extension of the Lee – Carter model to multiple populations.” *Annals of Actuarial Science*, 15(2), 346 – 366.

[13] Rothschild, M., and Stiglitz, J. E. (1976). “Equilibrium in competitive insurance markets: An essay on the economics of imperfect information.” *Quarterly Journal of Economics*, 90(4), 629 – 649.

[14] Yaari, M. E. (1965). “Uncertain lifetime, life insurance, and the theory of the consumer.” *Review of Economic Studies*, 32(2), 137 – 150.

[15] 一般社団法人生命保険協会（2022）「生命保険の引受・支払実務における遺伝情報の取扱につきまして」. https://www.seiho.or.jp/info/news/2022/20220527.pdf.

[16] 大倉 真人（2002）「リスク細分型保険は本当に望ましいか？」『経営と経済』82（3），79 – 94. https://nagasaki-u.repo.nii.ac.jp/record/27662/files/Keizai82_3_79.pdf.

[17] 公益社団法人日本アクチュアリー会（2018）「標準生命表 2018 の作成過程」. https://www.actuaries.jp/lib/standard-life-table/pdf/seimeihyo2018-katei.pdf.

[18] 二見 隆（1992）『生命保険数学』. https://www.actuaries.jp/examin/textbook/.

[19] 舩津 浩司（2014）「給付反対給付均等原則の法的再定位」『生命保険論集』189，99 – 126. https://www.jili.or.jp/files/workshop/search/D_189_4.pdf.

[20] 堀井 拓也（2016）「保険法 2 条 1 号の『保険契約』に関する一考察」『保険学雑誌』第 634 号，1 – 34 頁. https://doi.org/10.5609/jsis.2016.634_1.

世論調査と統計科学

公益財団法人日本世論調査協会
会長 鈴木督久

１．日本における世論調査の胎動

日本の世論調査は第二次世界大戦後に始まった。戦前にも新聞社による世論調査が実施されていたが、現在の世論調査の形式と内容は明確な起源を持ち、それが 2025 年現在まで約 80 年間、継続している。

1945 年 8 月 14 日にポツダム宣言を日本が正式に受諾して連合国に降伏すると、GHQ（連合国最高司令官総司令部）が日本の占領（1945～1952 年）を開始した。世論調査に関する政策は同年 9 月 22 日に設置された CIE（民間情報教育局）が担当し、下記のような方針が書かれていた。ここから日本の世論調査は出発した。

> d. 以下のような世論調査を促進し、その実行を指導することが必須である
> (1) 占領軍並びに復興に対する民衆の反応について事実に基づいた情報を最高司令官に提供し続けること
> (2) 政策並びに計画を連続的に作成し集成するための信頼できる基盤を確保すること
> ［Public Opinion Survey - Basic Plans. 翻訳は川島（1995）による］

CIE のミッションのうち世論調査は一部であり、大きな活動としては教育改革や広報戦略などであった。GHQ は新しい政策を実行する際に、新聞等を使ってその内容を日本国民に向かって説明した。CIE つまり Civil Information and Education の Information は情報と訳すことになっているが、広報の意味を多く含んでいる。

大きな影響を戦後日本に与えた教育政策では、文部省に改革を推進させるために、米国教育使節団を招いて政策提言となる報告書を作成した。この教育使節団報告書のうち、唯一の実行されなかった政策は「第 2 章 国語改革」である。それは民主主義の促進を理由とした、日本語表記に関する漢字廃止・ローマ字化の政策であった。CIE は新聞記事の検閲機関でもあった。一部の検閲官には、漢字よりローマ字であれば検閲業務が楽になる、という動機があったが、「日本文化を尊重せよ」という米国使節団と対立したのである。

ちなみに国語改革の妥協案は簡略化で落ち着き、当用漢字が制定された。この政策により輿論調査の輿が新聞記事で使えなくなり、朝日新聞社と毎日新聞社の担当者が「輿の代わりに世を使おう」と合意して世論調査（よろんちょうさ、と読む）が表記として定着した。輿論と世論の違いには、それ以上の深い意味はない。現在でも世論を「せろん」と読む人がいることは仕方ないであろう。仮名の「せ」は「世」が元である。歴史的な長さが違う。日本語なら「よ」だが、漢語なら「せ」と読むのが常識である。「世の中」という日本語は「よ

のなか」であって「せのなか」ではない。漢語の「世界」の世は「せ」と音読みして当然であるから「世論」を「せろん」と読むことは、教養としても自然である。中国語において輿は yu であり、世は shi であるから、漢字を輸入（呉音時代にせと漢音時代にせよ）した日本で、輿（よ）、世（せ）と読むことは当然であった。「よろん」と読ませたい新聞社の側に無理があり、ＧＨＱの占領政策に責任の源がある。占領時代の新聞社がなし得た対応としては、古代中国語ではなく現代中国語の「民意測験」を参考にして、輿論調査を非使用としたうえで、民意調査か民意測定に変更することであった。英語の public opinion research の訳語が民意調査であって不都合はなかったであろう。

世論調査は日本の民主化という大義が強調されてきたが、当時は占領下の日本民衆の本心を正しく把握したいという、直接的な動機があったことが窺われる。実際、GHQ は東京裁判に関する世論調査を実施することを企画して、千葉県での試験調査まで進めたが最終的には見送った。中止理由などを記録した GHQ 側の文書は、今のところ見当たらない。

CIE 局長にダイク（Kermit Reed Dyke）が就任すると、彼は世論調査を担当する部署を作ることを内閣情報局に勧め、日本側は輿論調査課を設置した。組織や名称に変遷はあるが、実質的な業務内容は現在の内閣府大臣官房広報室まで継続している。輿論調査課は意欲的に体制構築に着手したが、1946 年 6 月にニュージェント（Donald R. Nugent）が二代目局長として交代するや否や、彼は政府による世論調査を禁止した。表面的には技術・体制等が未熟だという理由が、輿論調査課に伝えられたが、ニュージェントは、政府ではなく民間で世論調査を実施すべきだと考えていた。戦前、ニュージェントは日本に住んでいた。二人の局長の異なる方針に対応したことで、内閣府と報道機関の官民それぞれで世論調査が実施されている現在の日本の姿が形成された。この経緯や背景については鈴木（2021）などを参照されたい。

CIE でも世論調査を担当する部署を作った。この組織も若干の変遷があるが、最終的な部署名は POSR（世論社会調査部）で、文化人類学者や社会学者などの言語将校が POSR のスタッフとなって、日本における世論調査の育成にあたった。彼らはまだ 30 歳代の理想を目指す若き研究者で、実証的研究の手法として調査法を使う分野にいた。帰国後は大学教授となっている。POSR については吉田（1994）なども参照されたい。

輿論調査課と POSR は、産官学から世論調査関係者を集めて輿論調査協議会を設置し、1947 年 3 月には米国から調査法・統計学の専門家を招いて研修会を開いた。場所は首相官邸広間で、通訳には鶴見和子（丸山真男らと「思想の科学」創刊）もいた。その後、輿論調査課は協議会を発展的に解消させ、その活動を民間で引き継ぐことを推奨した。それが現在の日本世論調査協会である。報道機関と研究者が主な構成員であった。この民間移転は、ニュージェントの意向を反映させる結果となっている。

POSR は勉強会などを実施したが、実際の調査を POSR が実施するには新聞社などの全国組織の活用（支局が調査員網となった）を必要とした。理論的側面では統計数理研究所に協力を求めた。その中では水野坦が CIE と兼務して、標本設計を中心に世論調査の確立に大きな貢献を果たした。統計数理研究所の水野研究室にいた西平重喜によれば、水野はほと

んど CIE に常駐していたという。もう一人は林知己夫であった。林は世論調査には関与しなかったが、CIE が企画した Literacy Survey（日本人の読み書き能力調査）のプロジェクトで標本設計を担当した。この調査結果報告書は戦後日本の標本設計・調査計画の手本となった。

水野は 1950 年に 100 日ほど訪米し、いくつかの大学で講義をした。占領下なので GHQ が水野の貢献を評価して、便宜を提供した渡航であろう。Kim（2023）はこれに関連して、Iowa State University で水野の講義を聞いたホービッツとトンプソンが、有名な HT 推定量（非復元抽出法における不偏推定量）に関する論文を書いたことに言及している。

> “the seminal paper of Horvitz and Thompson (1952) was published when the authors were graduate students at ISU and they were influenced by the lectures from Mizuno who visited from Japan.”

Horvitz and Thompson (1952) の刊行を巡っては、水野との間でトラブルがあったことが関係者の間で知られている。本件は明文化されていない挿話に過ぎないが、民主主義も科学的方法も、なにもかも米国に教わったかのような印象に対する反証のひとつであろう。

理論的側面だけではない。実践的にも日本は CIE が求めた理想的な世論調査を実現した。少なくとも米国では実践できない理想的な調査を日本は達成した。当時の米国では標本設計において非確率抽出 (有意抽出) 法による割当標本 (quota sampling) が主流であったが、日本は確率標本（probability sampling）による調査を確立したのである。

非確率標本よりも確率標本が、母集団をよく代表する標本であることは、Neyman(1934) で理論的に決着していたが、Gallup など著名な世論調査機関は、実際には割当標本を採用していた。1948 年の米国大統領選挙の予測では、主要な調査機関が予測に失敗したが、いずれも割当標本であった。国土の広さ、枠母集団となる名簿整備など諸条件を考えてみると、米国においては確率標本による世論調査を、全国規模で定例的に実施することは、現実には困難であろう。

米国の 1948 年大統領選挙での予測失敗は社会的にも大きな問題となり、ウイルクス（Samuel S. Wilks）など統計学者らが検証作業にあたった。Public Opinion Quarterly に掲載された報告書 (1948) では、標本抽出法に関して確率標本が推奨された。失敗の原因は複数列挙されているが、割当標本で予測したことが失敗の一因だと認定したのである。その点において比較すれば、日本は最初から理論的には理想的な確率標本調査を適用し、その実現に成功したのである。皮肉な見方をすれば、米国は自国ではできもしない理想を日本に教え、12 歳の少年（MacArthur：上院軍事外交合同委員会）がやり遂げたことになる。

日本で理想的な標本調査を実現できたのは、関係者の努力による。統計数理研究所の理論的貢献、報道機関各社の熱心な学習、GHQ／CIE の支援が総合的に結実した。この他に指摘すべき要因には枠母集団の充実がある。明治時代からの戸籍法の歴史的蓄積があり、配給台帳から住民基本台帳へと、ほぼすべての国民が義務として登録されている。しかも全国で

統一された基準と形式で整備され、標本抽出のために容易に閲覧できた。

諸外国では必ずしも、登録制度がなく、地方自治体によっても異なっている国もある。日本では住居は比較的固定的であり、共同体意識と行政への信頼感も高く、調査にも協力的であった。もちろん、現在の日本はそうではない。それは調査環境を悪化させている一因でもある。住民基本台帳は閲覧できなくなった。行政・報道機関への信頼も次第に低下しているだろう。特殊詐欺も増えており、見知らぬ人や電話への警戒感が高まれば調査へも非協力的になる。単身世帯が増加し生活時間も多様化しているし、マンションのセキュリティーも強化されたことで調査員が調査対象者と接触できる機会は減っている。

２．戦後 80 年間の展開

日本の世論調査は、GHQ／CIE だけでなく日本政府の世論調査部門、民間の報道機関、統計数理研究所など関係者の協力と集中的な努力で戦後数年のうちに完成した。政府世論調査と民間世論調査は、出発時点では同じ方法であったが、報道機関の世論調査が前半 40 年間と後半 40 年間で大きく変化した。2025 年時点で、今も同じ方法を堅持している唯一の例外は時事通信社（中央調査社）だけである。

表 1　報道機関の世論調査の変遷

年代区分	標本測定法	標本抽出法	目標母集団	枠母集団
1945－1984	面接聴取法	層化多段無作為抽出法	有権者	住民基本台帳
1985－2024	電話聴取法	RDD（三段無作為抽出法）	有権者	電話契約世帯
2025－????	？	？	有権者	？

注：面接聴取法から電話聴取法に変更した年代区分は各社で数年の幅がある

表 1 の年代区分は概略的で、ある日に突如として変更されたわけではない。戦後 80 年間をちょうど半分の 40 年で区切っているが、各社ごとに相違があり数年の幅がある。各社ともに 1980 年代に入ると、電話聴取法に取り組んでいるが、最初に定例世論調査を電話調査で開始した報道機関は日本経済新聞社（1987 年）であった。各社ともそれ以前から選挙情勢調査などの機会に、実験的に電話聴取法を実施しながら研究していた。また、標本抽出法も初期には住民基本台帳、電話番号簿などから始まり、現在のような RDD（Random Digit Dialing）を採用したのは、1997 年の毎日新聞社が最初であり，2001 年から朝日新聞社と共同通信社、2002 年に日本経済新聞社が続いた。NHK が 2004 年、読売新聞社が 2008 年で現在に至る。年代区分はそのような転換期間を持ちつつ、便宜的に 1 点（1985 年）を区切りとしたものである。報道機関の標本設計に関しては、「社会と調査」13 号（2014）で「選挙調査の最前線」、20 号（2018）で「メディアが実施する調査の変遷」の特集が組まれた。

世論調査の議論をする場合には注意すべき前提がある。標本抽出法（sampling）と標本測定法（measurement）を区別して議論をすることである。素朴に述べれば、誰に聞くか（抽

出)、どう聞くか（測定）ということである。世論調査の課題と未来に関する、最近のいくつかの議論をみる限りでは、両者が混同されているために議論が効果的でない印象がある。なお、標本測定法という用語はあまり使われないが、本稿の意図を考慮して採用した。一般には調査方法が使われるが、「**調査**対象者測定**方法**」の省略形とは認識されず、標本設計の広い範囲を含む誤解が想定されるので、端的に「測定」の意図を反映するように標本測定法とした。

両者は無関係ではないし、むしろ密接な関係があるというべきだが、区別が必要となる背景がある。標本抽出法は理論的な支柱であり、標本測定法は実践的な現実性を支えている。ある時代と社会環境のもとで最適な標本測定法を考慮し、その実現のための標本抽出法を考案する、という関係になることが多かったと思われる。

3.　戦後の前半 40 年間（面接聴取法の時代）

1940 年代の日本においては、調査員が調査対象者の家を訪問して、調査への協力を依頼してから面接を実施し、調査票に回答結果を記入する方法が適切であった。通信手段としては郵便しか使えなかったが、郵便では調査内容の説明がすべての調査対象者に伝わるか懸念があった。訓練された調査員が調査の意義を含めて協力依頼することが丁寧な態度であり、見知らぬ調査員を玄関に迎えることに抵抗も少ない時代であった。その結果として「調査員による訪問面接聴取法」が最適な標本測定法として選択された。

標本抽出法に影響する現実的問題は、調査員は自動車ではなく徒歩で活動する前提なので、担当できる地理的範囲が小さいことである。一人の調査員が調査期間内に回収できる件数を計算すれば、標本規模から逆算して、必要な調査員の人数が決まる。標本規模は統計科学的に決める。多くの場合は 95%信頼区間の公式を使い、目標精度を定めて標本規模を決める。支持率（割合）の精度として報道機関が採用しているのは、95%信頼区間を 1 / sqrt(n) と保守的に見積もって、±2〜3%ポイント程度を目標としている。内閣支持率などを整数に丸めて公表している統計科学的な理由は、百分率の小数点以下は有効桁ではなく誤差だからである。

ここで標本規模とは sample size の訳語である。多くの教科書では標本の大きさ、標本サイズ、サンプル数、標本数などの訳語が使われている。このうちサンプル数、標本数は統計学の用語としては間違いである。これは number of samples の意味になってしまう。標本規模という新しい用語を使うことで、サンプル数という誤訳を駆逐する効果も期待できるだろう。JIS 規格では「標本の大きさ」であるが、筆者の好みで「用語は漢語で」の方針に馴染みがある。漢語と和語を組み合わせると「大きさ比例確率抽出」のようになるが、いかにも不自然ではないか。あるいは「大きさに比例する確率のサンプリング」のような訳例もあるが、直訳というよりもはや一文に近い。なぜか不明であるが、sample size（標本の大きさ）だけは、size を規模と訳さずに大きさと訳す例外の用語となっている。ただ、戦後最初の確率抽出を適用した「労働力調査」の報告書では標本規模と訳しており、終戦直後には適訳が

あったのに自然淘汰されてしまった。余談だが、random sampling の場合は、任意抽出法が自然淘汰され、無作為抽出法が定訳となった。

結論として標本抽出法は「層化多段無作為抽出法」が選ばれた。標本規模は許容できる誤差の大きさを考慮すると 3000 人程度が全国世論調査で適切だとされた。全国規模の調査を実施するには多数の調査員が必要となる。全国に 1 人 1 票（調査対象者）で配置すれば、必要な調査員は 3000 人となるが、それはさすがに非効率なので、1 人 10 票とすれば 300 人の調査員体制となる。1 人 2 地点を担当すれば 150 人体制である。

そこで多段抽出法が採用された。全国から第一次抽出単位として 300 地点を選び、最終抽出単位の有権者は、各地点で調査員が担当する仕組みである。統計理論的には、多段抽出法は一段抽出法よりも標準誤差が大きくなるので、層化抽出法を組み合わせることが考案された。どの程度の悪化となるかは前田・中村（2000）を参照されたい。適切な層化をすれば標準誤差を小さくできるので、多段抽出法の欠点を補うことが期待される。それが層化多段抽出法で決着した理論的および実践的な理由である。

なお、無作為抽出法は確率抽出法と呼ぶ方が適切だが、歴史的に無作為抽出法と呼ばれてきたので、本稿でも文脈によっては習慣に従いつつ、原則としては確率抽出法とする。この用語問題については土屋（2018）や日本統計学会編（2023）のコラムも参照されたい。実際に地点を無作為抽出する場合には系統抽出法が実践的に有効である。

世論調査に適用された標本抽出法は複数の抽出法の組み合わせであり、それらを列挙すれば以下のようになる。個々の抽出法に関する具体的な手順の詳細は省略する。数理的性質についは多くの統計学のテキストに解説があるが、たとえば土屋（2009）を参照されたい。

- 多段抽出法
- 層化抽出法（層別抽出法）
- 規模比例確率抽出法（確率比例抽出法）
- 系統抽出法（等間隔抽出法）

具体的な事例として「日本人の国民性調査」の第 8 次調査（1988 年）以降で採用されている層化三段無作為抽出法の手順を確認しよう。詳細は、国民性調査委員会（1992, 1999）や前田・中村（2000）を参照されたい。

表 2 第 10 次 国民性調査（1998 年）の標本設計

母集団	全国の有権者
枠母集団	有権者名簿
計画標本規模	4,200（n）
計画地点数	300（m）
1 地点当り標本規模	14（n/m）
層の数	5（L）
層	区部／人口 20 万人以上の市／人口 20 万人未満の市／郡部／沖縄県（h）

・　層の規模に応じて計画標本を比例割当して、各層の標本規模（n_h）を決める
・　層の規模に比例させて各層の地点数（m_h）を割当する
・　各地点の標本規模を決める（平均的に 14 だが若干の端数処理が必要）

以上を踏まえて、以下の手順で標本抽出をする。
(1) 第一次抽出単位である市区町村を規模比例確率抽出する
(2) 第二次抽出単位である投票区を、抽出された市区町村から規模比例確率抽出する
(3) 第三次抽出単位（最終単位）である有権者を有権者名簿から系統抽出する

この三段抽出の手順は簡単に記述したが、概念的に分かりやすくするためである。実際は巧妙な「実践的工夫」をしている。それは 2 点に集約される。まず、市区町村を層化効果が発揮されるように並べること。次に、仮想的な地点（調査単位集団）を導入したことである。

結果的に上述のとおりに三段抽出を実現しているが、「実践的工夫」は統計学の教科書には書かれていない。たとえば、仮想的な地点（調査単位集団）という考え方は、林・村山（1964）などに手順の説明があるが、統計理論を踏まえたうえで、約 1 億人の有権者が並んでいる状態を想像できないと理解が難しい。しかし実に具体的で実践的な作業手順であり、しかも理論的にも満たされている。実際に台帳をめくって標本抽出をした経験がなければ思いつかない。あるいは、標本抽出を深く考え、等しい確率で抽出される結果を得る手段に想像力を発揮できるほどの、本質的理解が必要であろう。多くの統計学の教科書は理論の知識しか書かれていないか、「壺から玉を取り出す」ような比喩しかないので、教科書としては立派に役割を果たしているが、現場では役に立たない。市区町村や投票区などの行政的な境界区分は、いったん捨象して本質を抽象すれば理解しやすいかと思われる。

もう一つの工夫である「市区町村を並べること」は驚くべき発想である。水野や林の時代は数十にも達する多くの層を作っていたが、実際面では不都合もあった。そこで 5 層に減らすが、多数の層化をした場合と同様の効果を発揮するように並べてから系統抽出するのである。ただし、人口統計データベースとコンピュータがなければソートの手間がたいへんである。具体的な手順は国民性調査委員会(1992)などを参照されたい。

ヘーゲルはどこかで「理性的なものは現実的であり、現実的なものは理性的である」と述べたが、統計学と調査法も「理論的なものは実践的であり、実践的なものは理論的である」と言ってみたくなる。

4.　戦後の後半 40 年間（電話聴取法の時代）

電話調査は米国が先行していた。米国で電話が普及した 1960 年代以降、まず市場調査から始まり、やがて理論的研究としても多くの論文が発表され、公的分野や学術分野にも電話調査が承認された。日本では 1980 年代から電話調査の時代となったが、これは日本で電話の普及がピークに達した時期である。つまり電話世帯という枠母集団のカバレッジ誤差が

小さくなった時に、はじめて電話調査のリアリティーが生まれる。

外的環境が整ったという要因のほか、報道機関の内的動機も重なった。世論調査とは異なる側面もあるが、同じ方法で実施していた選挙情勢調査（選挙予測調査）の巨大化である。衆院に小選挙区比例代表並立制が導入され、1996 年総選挙から適用された。中選挙区制では 132 選挙区であったが、一気に 300 選挙区と倍以上に増えたことで、調査予算の確保が難しい状況になったと同時に、実査面でも調査員の確保は難問となった。

報道機関の電話調査は、まず選挙情勢調査から始まったが、続いて世論調査に関しても、小泉純一郎内閣（2001～2006 年）の登場で、世論調査が頻繁に求められるようになった。速報性も重視され、それが電話調査の特性と適合した。現在、報道機関の定例世論調査は月次で実施されているが、その頻度と迅速性は電話調査だから可能であった。

標本抽出法に関しては、最初は調査員調査の時代と同様に、住民基本台帳からの確率標本抽出や、電話番号簿からの系統抽出法が利用された。しかし電話帳への掲載率が年々減少して、枠母集団のカバレッジ誤差が無視できない程に増大した。RDD sampling への移行が不可避になった。報道機関の世論調査が RDD に移行した時期は、2000 年前後であったが、それは小泉純一郎内閣と重なっている。

RDD sampling に関しては、Waksberg（1978）が大きな影響を与えた、日本における報道機関の RDD もここが出発点となった。この重要な論文は Warren Mitofsky が CBS で 1970 年までに実際に運用していた二段無作為抽出法に理論的根拠を与えた。電話調査において確率標本抽出が可能であるという合意形成がされた画期的な論文であった。Mitofsky が Waksberg に理論的検討を依頼したもので、Mitofsky が米国センサス局で仕事を始めた時の上司が Waksberg であった。

5. Mitofsky - Waksberg 法

電話調査でも目標母集団は有権者である。枠母集団に関しては、大別すると下記の二種類の戦略がある。枠母集団の要素を列挙した枠（frame）を標本抽出枠ともいう

- working exchange frame（working exchange design）
- list-assisted frame（list-assisted design）

稼働局番フレーム（working exchange frame）は、稼働中のすべての電話番号を枠母集団とし、この枠を使った標本設計を稼働局番法（working exchange design）という。日本の場合、総務省が稼働局番を公表しているので、容易に完全なフレームを作成できる。

名簿準拠フレーム（list-assisted frame）は電話番号簿を中心として各種名簿（電話帳以外にも情報源があれば加える）から構築した電話番号データベースから、世帯番号を基準数（たとえば 1 件）以上含むバンク（bank）を列挙する。このフレームを使った標本設計を名簿準拠法という。

目標母集団に対するカバレッジ誤差は名簿準拠フレームのほうが大きい。世帯番号非掲載（基準数未達）のバンクが排除される切捨て（truncated）が発生するからである。切り捨

てたバンクには世帯番号が存在する可能性があり、名簿に掲載していないだけかも知れないからである。しかし世帯ヒット率は名簿準拠フレームのほうが高くなるので、実査は稼働局番法より効率的になる。

ここでバンクとは、10 桁の電話番号に関して上位桁が等しい番号の集合である。クラスターということもある。百位バンクとは 2 桁バンクであり、上位 8 桁が等しい 100 個の番号（00〜99）を持つ。千位バンクは 1,000 個の番号を持つ。万位バンクは局番であり 10,000 個の番号（0000〜9999）を持つ。世帯番号を含むバンクを有効バンク、含まないバンクを無効バンクという。

Mitofsky - Waksberg 法は、稼働局番フレームを使った多段無作為抽出法である。以下の手順で標本とする電話番号を抽出する。

ちなみに、一般的な RDD の説明文の中には「コンピュータで電話番号を無作為に発生させる」と書いているものがあるが、あくまでも母集団（標本抽出枠である番号の集合）からの抽出であって発生と考えないほうが、標本抽出法に対する適切な理解につながる。機械がランダムに番号を「作り出して」いるかのような印象を与え（現象的にはその通りだが）、統計理論が伴っていないような誤解を与えるからである。

(1) 第一次抽出（PSU）

百位バンクを単純無作為抽出し、各バンクで二桁の乱数を 1 つ発生させて完全な電話番号を 1 個作る。そこに電話をかけて、世帯用番号だったらそのバンクを保持し、そうでなければそのバンクを捨てる。結果として「世帯用番号の件数」に比例する確率で m 個の百位バンクが抽出される。世帯番号を含まないバンクは決して抽出されない。

(2) 第二次抽出（SSU）

各百位バンクで二桁の乱数を発生させて完全な電話番号を作り「一定数 k の世帯用番号」を得るまで続ける（たとえば k=5 とする）。結果として、抽出率はバンクの世帯用番号数に逆比例する。

(3) 第一段と第二段の抽出確率

両段階の抽出確率の積は各 bank において等しく、各世帯が等しい確率で抽出され、合計で mk 個の番号を得る。なお、第一段階と第二段階の抽出作業は独立に実施してもよい。

(4) 第三次抽出（TSU）

各世帯の有権者から 1 人を単純無作為抽出して調査対象者とする。有権者個人の抽出確率を等しくするために、データを集計する際に、有権者数 a と世帯回線数 b を重み (a/b) とする。

Waksberg（1978）の画期性は、第一段および第二段の抽出確率（bank にいくつの世帯番号があるか）を事前に知る必要がなく、にもかかわらず、すべての世帯番号を等確率で抽出できることを理論的に証明したことである。PSU は規模比例確率抽出になっているが、その規模は実際に PSU に存在する正確な世帯番号数であり、過去の国勢調査データなどの推定値ではない。

Mitofsky - Waksberg 法は、電話世帯を完全にカバーし、第二段の世帯ヒット率が大幅に高まり（20%→60%）、RDD の欠点を克服した。第一次抽出において、世帯番号が存在しないバンクは、決して抽出されないことが重要で、これによって効率化を実現したのである。

もちろん短所もある。標本抽出と標本測定（実査）が、絡み合っているので管理が複雑になる。前のプロセスが完了しないと、次のプロセスに進めない。この順序性の制約で時間がかかるが、第一次抽出と第二次抽出は同日に実施する必要はない。なお、非常に低い確率とはいえ、第一段で世帯番号が k 個未満のバンクが選ばれると、100 個かけ終わっても k 個に達しない可能性がある。しかし、理論的には RDD サンプリング法はここで完成を見た。

ところが、日本では電話調査に対して研究者から批判を浴びた。十分な研究もしないまま報道機関が電話調査を始めたという心情的な批判もあった。報道機関は論文を書くことは仕事ではないから、社内の研究結果は論文として刊行されるわけではない。確かに戦後すぐに水野坦らが集積した研究の分量に比べれば、日本の電話調査の研究は少ない。

林知己夫編（2002）『社会調査ハンドブック』では RDD 電話調査について、以下のように説明している。

> 選挙人名簿から対象者を特定する方法では、「母集団」が有権者全体になるが、RDD では明確に「母集団」を規定できない点である。RDD で作成した番号には、世帯用の番号だけでなく、事業用や使われていない番号も含まれている。調査時点では、世帯用番号なのかそうでないのかを判別できない番号も多いからだ。

この記述は正しいであろうか。RDD 電話調査の目標母集団は有権者世帯であり、調査員調査と同じである。枠母集団は電話世帯である。明確である。世帯電話番号（適格番号）と非適格番号が混在しているのは当然であり、Waksberg（1978）で見たように理論的に問題ではない。実践的に効率性の問題があることが RDD の課題に過ぎない。

RDD 電話調査に対する理解は必ずしも浸透していない証左である。関連して「回収率が曖昧である」という批判もある。回収率には複数の定義がある。表 3 は RDD 電話調査の結果のまとめで、報道機関は計画標本番号がどのような結果で終了したか（disposition）をこのように記録している。計画標本規模は 18,000 件から始まり、まず非使用番号を除去した 6,428 件に電話をかけて、有権者のいる世帯か否かを確認する。その結果、不適格番号（事業所等）が 1,624 件、対話できなかった番号が 886 件であった。対話できたものの適格番号か否かを確認できなかった（ガチャ切り等）番号が 137 件――ということを示している。ちなみに調査員調査でも留守のために確認できない対象者は残る。その点では同じである。

回収率に関しては 3 種類が算出されている。もっとも高い回収率は 63%であり、この分母は「有権者がいることを確認できた番号」である。現在、報道機関が示している回収率はこの定義による。要するに「もっとも甘い定義」である。次に、世帯電話のようだが、有権者がいるか質問した段階で回答拒否された 576 件を分母に加えた 53%が示されている。最後に世帯か事業所が回答を得られなかった 137 件を分母に加えた 51%である。ここまでは

世帯番号である可能性はかなり高いであろうと推察される。もっとも甘い（高くなる）定義を使っていることは間違いないので、そこを批判の論点にするのであれば問題は無い。しかし、これで電話調査では「明確に母集団を規定できない」と解説することは不当ではないかと思える。

報道機関としては内部ではこのようなデータが保存されている。調査概要は記事の中では詳細まで報告されない習慣がある。調査結果の解釈が報道記事の中心である。記事の中では書かないにせよ、WEB サイトにおいて公開するなどの対応は可能であろう。

表 3 RDD 電話調査の回収率

<table>
<tr><td colspan="2">抽出番号数（18,000）</td><td>非使用番号除去後</td><td>6,428</td><td colspan="3">100%</td></tr>
<tr><td colspan="2">非適格</td><td>事業所，ＦＡＸなど</td><td>1,624</td><td colspan="3" rowspan="2"></td></tr>
<tr><td colspan="2">不対話</td><td>呼出音，話中音・留守電</td><td>886</td></tr>
<tr><td colspan="2">不明</td><td>世帯か事業所かも未確認</td><td>137</td><td colspan="2" rowspan="4"></td><td rowspan="4">3,918</td></tr>
<tr><td rowspan="3">世帯</td><td>有権者がいるか確認できない世帯</td><td rowspan="2">拒否</td><td>576</td><td rowspan="3">3,781</td></tr>
<tr><td rowspan="2">有権者のいることを確認できた世帯</td><td>1,195</td><td rowspan="2">3,205</td></tr>
<tr><td>回収</td><td>2,010</td></tr>
<tr><td></td><td colspan="2">調査期間：2003 年 7 月 31 日～8 月 3 日</td><td>回収率</td><td>63%</td><td>53%</td><td>51%</td></tr>
</table>

6. 日本における RDD 電話調査の実装

報道機関各社は米国の研究成果を参考に、独自の実験研究を積み重ねて実施体制を構築した。各社で細部は異なることは、公表された範囲の論文等で分かるが、大きな相違は名簿準拠法か稼働局番法か、にある。Waksberg（1978）以降の米国での状況の変遷は、島田（2003）を参照されたい。

朝日新聞社に関しては、松田（2002）によれば名簿準拠法である。バンクを採用する基準数は明らかにされていないが、米国での例では 1 以上が多い。

日本経済新聞社に関しては、島田（2005）が示すように稼働局番法である。しかも多段抽出法ではなく一段階の単純無作為抽出（系統抽出）法である。使用されていない番号を、呼出音を出さずに信号検出するシステムが日本では利用可能であり、これによって効率化が達成できる。実験結果によれは、単純無作為抽出であっても Mitofsky – Waksberg 法に匹敵する効率性であり、実査上は十分であった。もちろん名簿準拠法はさらに効率的であるが、

カバレッジ誤差が無く、名簿準拠フレーム（電話番号データベース）を維持するコストも不要であり、総務省が公表する最新の、すべての可能な電話番号（all possible numbers）を枠母集団にできることを重視した判断であった。

朝日や日経以外の報道機関においても、Waksberg（1978）の理論的達成を起点にしつつも、実査面、技術面などの観点から修正を加えているであろう。また電話調査の環境変化の速さに対応する必要にも迫られる。たとえば、名簿準拠法において切捨て（truncated）による誤差は小さいとの報告もあるが、その後の電話環境の変化にも対応できるかは疑問である。理論と効率的実践の両立を目指す Casady and Lepkowski（1991）のような非比例層化抽出法のアイデアも考案された。

しかし、もっと大きな環境変化は携帯電話の普及である。これは RDD サンプリングの細部の相違など問題にならないくらい大きな課題となった。固定電話を契約せずに、携帯電話しか持たない世帯が増加の一途をたどり、カバレッジ誤差の増大は無視できなくなった。この課題は各社共通であったことから、日本世論調査協会の有志会員で共同研究が実施され、2015 年に報告書が出された。その後、各社は相次いで携帯電話を調査対象に加えていった。

固定電話だけを前提とした Waksberg（1978）は、わずか 40 年程度で早くも現実的課題に直面した。携帯電話を枠母集団に加える場合、異なる母集団からの抽出と考える dual frame h 法と、固定電話も携帯電話も合わせてひとつ電話番号フレームと考える single frame 法が考えられるが、理論的にも実践的にも決定的な解決には至っていない。福田（2017）によれば読売新聞社は dual frame 法を採用した。槙（2017）によれば日本経済新聞社は single frame 法を採用した。

7. 現状（2025 年）の課題と未来

ここまで標本抽出法について記述してきた。ここに統計科学が中心的に使われ、科学的な世論調査であると主張できる根拠があるからである。標本測定法については記述しなかったが、これは科学的ではないということを意味しないし、重要ではないということではない。むしろ世論調査や社会調査では極めて重要である。

戦後前半 40 年においても、調査員の教育研修、調査票（質問票）の設計などはデータの品質管理のうえで重視されてきた。調査結果に直接反映する要因であり、研究成果も非常に多い。

戦後後半 40 年における電話調査でも、オペレータの教育研修は大きな部分を占める。調査票（質問文）も同様である。固定電話と携帯電話では、調査結果における回収標本の属性（年代や性別）分布に明確な差異があることも知られている。実査会場の状況も異なる。固定電話の場合は在宅時間にはつながるが、それ以外の時間帯は調査の進捗が遅くなる。一方、携帯電話の場合は時間帯と関係なくつながり、調査が進む。決して、標本測定法を軽視するものではない。

調査員調査から電話調査への移行は、調査環境の変化に迫られた側面もあった。面接聴

取法に関しては、回収率の著しい低下である。図 1 は日本を代表する社会調査である「日本人の国民性調査」（統計数理研究所）と「日本人の意識調査」（NHK）の回収率である。戦後当初は 80%もあった回収率は、60 年間で 50%まで低下した。

電話調査でも回収率が高いわけではない。報道機関にとっては速報性の重視、予算の制約という要因が強かった。回収率だけを問題にするのであれば、現在では郵送調査が最も高く、調査員調査を上回る。そのため、報道機関は年に数回ではあるが、郵送世論調査を併用して、電話調査の欠点を補う努力をしている。しかし、電話調査の回収率の低下は限界に近い。2 日間しか調査期間がないためだけではなく、見知らぬ電話に出る人は減少している。日本社会は調査への協力的態度そのものが縮小しており、方法の転換だけでは解決しない。調査環境の悪化は米国と同様の未来に向かっているかも知れない。もはや胸を張って「理想的な標本調査を実現した」とは主張できない時代になった。米国より、少しだけマシという程度であろう。米国における電話調査の回収率は 5%（！）だと聞いたことがある。現在はもっと低下しているだろう。さすがに日本はそこまで壊滅的ではないが、未来の姿かも知れない。戦後一貫して、米国の悪い側面に日本社会全体が追随したという事かも知れない。

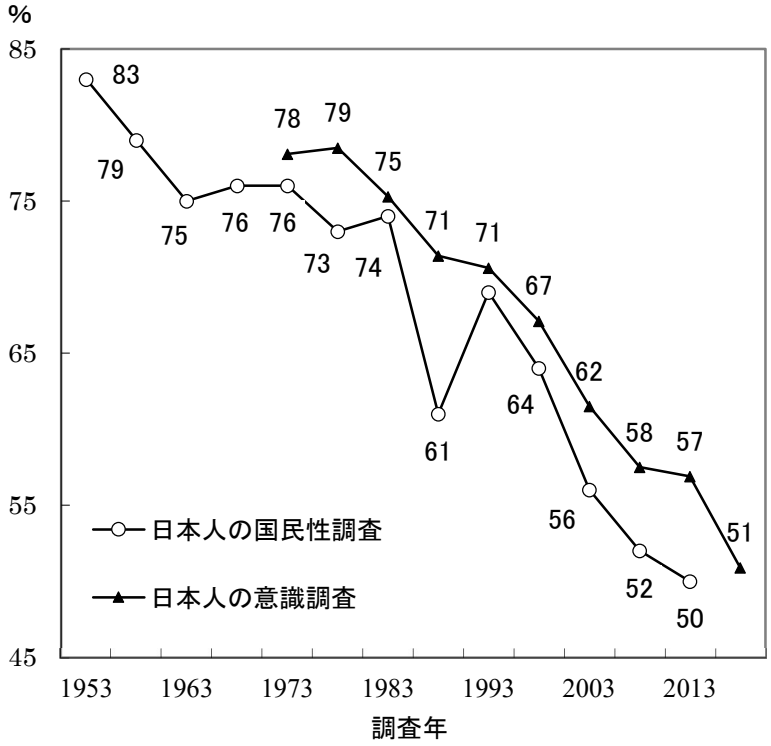


図 1　代表的な社会調査の回収率推移

戦後 80 年を振り返った。もしも今が限界点ならば、ここから将来を展望しなければな

らない。しばしば聞こえるのは「インターネットを利用せよ」「調査も DX 化せよ」などの声である。そうかも知れない。しかし、その議論において、冒頭で強調したように標本抽出法と標本測定法を区別して議論する必要があることを合意しておきたい。インターネットの利用はよいが、ほとんどの議論は標本測定法に関する内容になっているのではないか。

標本測定法としてインターネットを利用することはよいだろう。利点は非常に多い。しかし、そのような主張をする人から確率標本抽出法の提案はあるだろうか。もちろん、いくつかの方法はある。

第一に、有権者名簿（住金基本台帳）から確率標本を抽出する。これは昔と同じである。学術研究調査であれば適している。しかし報道機関の世論調査では迅速性に劣る。それをするなら郵送調査にすればよいかも知れない。むしろ郵送調査より手間と時間がかかるであろう。抽出した対象者に郵便を出して、メールアドレスを返送してもらうか、調査用サイトへアクセスしてもらう必要がある。本人確認をする方法も必要となる。それは技術的に可能であろうか。未回収の対象者には電話番号を調べて督促や疑義照会も必要になる。郵便かインターネットかという選択の相違に過ぎないが、インターネットで測定できるメリットもあるだろう。

第二に、民間調査機関のインターネット調査を利用する。安くて速い。しかし枠母集団はわずか数百万人のボランティア・モニターであり、目標母集団の有権者から大きな偏りがある。確率標本ではなく割当標本なので、統計科学的な推定ができない。割当標本（非確率標本）が確率標本に劣ることは、Neyman（1934）で理論的に決着している。この選択は 1920 年代へと 100 年ほど退化することを意味するのではないか。

第三に、インターネットならではの技術の応用である。一例として、Random Domain Intercept Technology (RDIT)がある。これは無作為抽出法にはなっているが、確率抽出法になっているだろうか。世論調査の目標母集団は有権者である。RDIT の枠母集団は調査期間中にインターネットにアクセスした人々のうち、URL を間違えてタイプした集団、ということになるだろうか。カバレッジ誤差は大きく、調査期間中にインターネットにアクセスしなかった人は含まれない。報道機関が「現在の世論だ」ち主張する根拠としては弱いであろう。

第四に、マイナンバーを利用する。戦後日本の世論調査の成功の原因は、住民基本台帳（あるいは配給台帳）の充実であった。標本抽出のために閲覧できた。マナンバーが住民基本台帳の DX 化であれば、マイナンバーを標本抽出に使えば、確率標本に対するインターネット調査は成功するだろう。しかし、住民基本台帳の閲覧は法律で禁止され、世論調査の目的では閲覧できない。かろうじて選挙人名簿は報道機関として閲覧できるので、同様に有権者のマイナンバーを世論調査の標本抽出に利用することが考えられる。しかし、そのためには法整備が必要であり、その前提として国民の合意が求められるだろう。技術的には、マイナンバーにメールをするためには、マイナンバーがメールアドレスになっており、国民全員が個人のメールアドレスを持ち、インターネットに毎日アクセスする社会

の実現が必要である。現状では非現実的であろう。

最後は、統計科学あるいは近代科学のパラダイムの外に出ることであろうか。もはや調査法によってデータを得ることを断念する。デジタル化社会の進展、人工知能（AI）の利用で、調査をしなくても有権者の意見・意識を把握できるのであれば、有権者が非協力的な調査法を続ける必要はない。具体的な方法は、いろいろ想像する以外にはよく分からない。従っていつ実現するのかも分からない。いつかはそのような社会が到来し、その未来からみると現在のことは歴史的に「調査の時代」であった、と記述されるかも知れない。

文献

Casady, R.J. and Lepkowski, J. M. (1993). Stratified telephone survey designs. Survey Methodology, June 1993 Vol.19. No.1.

Committee on Analysis of Pre-Election Polls and Forecasts of the Social Science Research Council. (1948). Report on the analysis of pre-election polls and forecasts. *Public Opinion Quarterly,* 12(4). 599-622.

GHQ/SCAP Records (1946-51). Public Opinion Survey - Basic Plans. Civil Information and Education Section, Public Opinion and Sociological Research Division. General Subject File, 1946-51, Box:5874, folder:11. 国立国会図書館請求記号 CIE(B) 07424-07426.

Hiroshi Midzuno (1950), An outline of the theory of sampling systems. Annals of the Institute of Statistical Mathematics (AISM),1, 149-156.

Kim, J. K. (2023). Survey Sampling History at Iowa State University. The Survey Statistician, vol 88, 51-57.

Neyman, Jerzy (1934). On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection, JRSS, Vol.97.

Waksberg, Joseph. 1978. Sampling Methods for Random Digit Dialing. Journal of the American Statistical Association 73:40–46.

川島高峰（1995）．戦後世論調査事始－占領軍の情報政策と日本政府の調査機関－.メディア史研究 第二号．ゆまに書房.

携帯 RDD 研究会：日本世論調査協会会員有志（2015）．携帯電話 RDD 実験調査結果のまとめ．日本世論調査協会.

国民性調査委員会（1992）．第 5 日本人の国民性－戦後昭和期総集．出光書店.

国民性調査委員会（1999）．国民性の研究 第 10 次全国調査―1998 年全国調査―．統計数理研究所研究リポート 83．統計数理研究所.

佐藤寧（2020）．1948 年アメリカ大統領選挙予想は何故失敗したか－当時の世論調査協会資料には何が書いてあるのか．よろん，126.

佐藤寧・槙純子（2008）．RDD サンプリングにおけるフレーム比較．行動計量学，35（2）．
島田喜郎（2003）．米国における RDD サンプリングの最近の動向．日本行動計量学会第 35 回大会発表論文抄録集．
島田喜郎（2004）．RDD サンプリング手法の比較研究．よろん，93．
島田喜郎（2004）．RDD サンプリングの理論と実践．マーケティング・リサーチャー．97．
島田喜郎（2005）．RDD サンプリングにおける稼働局番法の再評価．行動計量学，62．
鈴木督久（2009）．世論調査の最近の動向．社会と調査．3．
鈴木督久（2021）．世論調査の真実．日本経済新聞出版．
鈴木督久（2025）．日本におけるギャラップ伝説の終焉―1936 年の奇跡の軌跡―．In：政治空間における諸問題：有権者，政策，選挙．第 6 章．中央大学出版部．
土屋隆裕（2009）．概説 標本調査法．朝倉書店．
土屋隆裕（2018）．無作為抽出法と有意抽出法．社会と調査．20．
日本統計学会編（2023）．調査の実施とデータの分析．東京図書．
林知己夫編（2002）．社会調査ハンドブック．朝倉書店．
林知己夫・村山孝喜（1964）．市場調査の計画と実際．日刊工業新聞社．
福田昌史（2017）．固定電話と携帯電話を対象とした電話調査の導入と推定値の評価．行動計量学，44(1)．
前田忠彦・中村隆（2000）．近年 5 回の国民性調査の標本設計と標本精度について．統計数理，48(1)．
槙純子（2017）．シングルフレームによる固定電話・携帯電話併用式 RDD 調査．社会と調査．17．
松田映二（2002）．朝日新聞社の RDD 調査について．行動計量学，29(1)．
水野坦・林知己夫・佐藤良一郎（1951）．サンプリング調査法．朝倉書店．
吉野諒三（2002）研究の意義と実験調査結果の考察．In：内閣府大臣官房政府広報室（2002）．『平成 13 年度 世論調査に関する調査研究－世論調査のサンプリング方法についてー研究報告書』第 2 章．内閣府．
吉田潤（1994）．占領軍と日本の世論調査―ベネットの POSR 資料から―．NHK 放送文化調査研究年報，39．
読み書き能力調査委員会（1951）．日本人の読み書き能力．東京大学出版部．

おわりに

統計数理研究所が推進している統計エキスパート人材育成事業の一環として2025年4月〜6月にかけて「連続講義シリーズ:社会の中の統計科学」が企画された。本講演シリーズでは統計学・統計科学がよく利用されている社会における幾つかの重要な分野に絞って、統計学・統計科学が有効に利用されている側面と共に負の側面にも言及し、課題を正しく理解し、今後の統計科学の教育の現場において適切に対処する能力を養うことを目的に行われた。連続講義シリーズは日本の各分野を代表する専門家による講演からなり、各講演者は多忙な中にもかかわらず講義を実施していただいた。この連続講義シリーズの企画者としてはこの場を借りて改めて感謝する次第である。

日本における統計科学分野における統計エキスパート人材の必要性が叫ばれている中で日本における統計科学および応用の諸分野における統計エキスパート育成として妥当な内容は何か、今でもなお模索中である。今後の教育内容を改善する参考の為にも、各界の統計科学や応用諸分野などの関係者からの忌憚のないコメントを期待したい。

2025年10月
国友直人（連続講義シリーズ・コーディネーター）