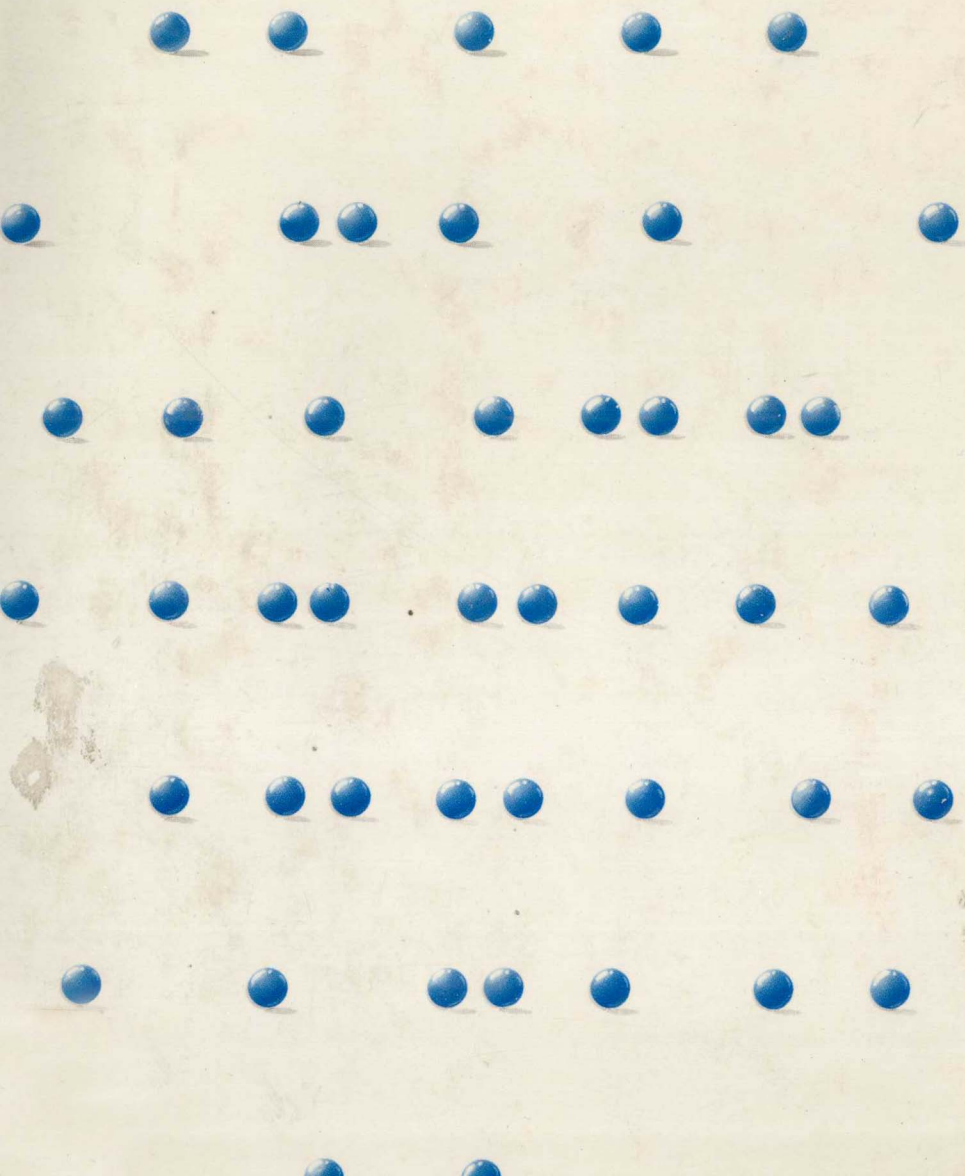


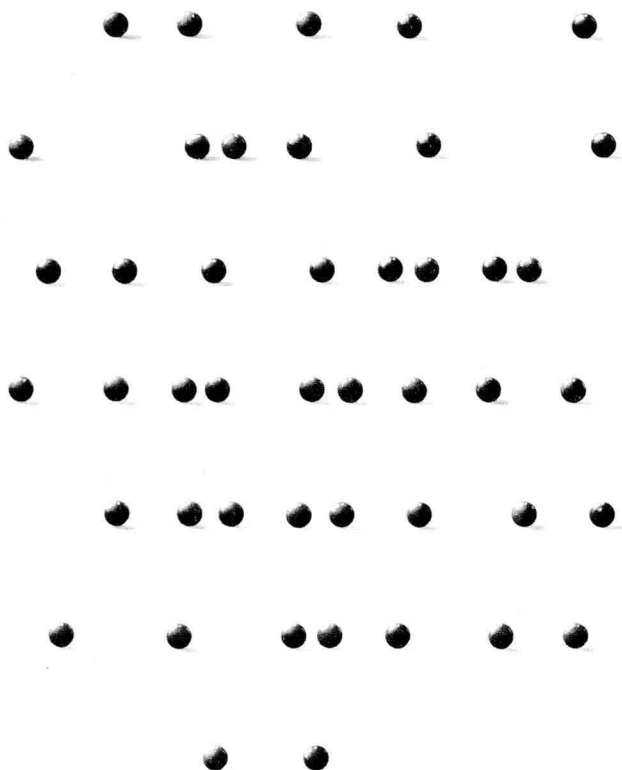
論調查本標解例

統計数理研究所 理学博士 鈴木義一郎 著



論調查本標解例

統計数理研究所 理学博士 鈴木義郎 著



著者略歴

鈴木義一郎

昭和12年9月19日山形県に生まれる。昭和37年東北大学理学部数学科修士課程修了。現在統計数理研究所第三研究部第二研究室長。理学博士

主要著書

例解統計入門(実教出版)

統計解析術—データを活かして使う法—(実教出版)

モデル解析術—O R 的考え方の基本原理—(実教出版)

データ解析術—“記述統計”のすすめ—(実教出版)

基本統計学(共著)(実教出版)

統計術指南(日本経済新聞社—日経新書)

統計解析法の原理(朝倉書店)

「比較」統計学のすすめ(講談社—ブルーバックス)

統計いろはがらた(東洋経済新報社)

「勝負」の統計学(講談社—ブルーバックス)

例解標本調査論

N D C 417

1981年10月30日 第1刷発行

定価 1400円

著者 鈴木義一郎

発行者 宇野豊蔵

印刷 株式会社廣濟堂

発行所 実教出版株式会社

東京都千代田区五番町五番地 下102

電話 東京(263)0111(大代表) 振替東京 4-183260

© G. SUZUKI 1981

3041—2220—3205

は じ め に

おかみさんが鍋の中にみそを入れ、よくかきまぜてからお玉ですくって塩梅をみる。これは、鍋全体のみそ汁の味とお玉ですくった一部分の味とが、一致することを期待しての行為に他ならない。

本屋に入れば、手にした本の頁をばらばらとめくりながら、それを購入して読むべきかどうか判断する。

初めて入る食堂で、店頭のケースの中のサンプルと値段とをみくらべてみる。ろう細工のトンカツは大きかったのに、目の前に出された代物のなんと小さいこと。

たった一週間から十日間程度の外国旅行をして、その国のほとんどのものを見聞きしたと錯覚している旅行者。

あるとき、ある人が自己の中のかくされた才能に気づいたりするのも、自己認識すら全体的にはよく観察できないことの証左でもある。

このように、われわれの認識の大方のケースで、全体のある一部分だけに依存して判断していることが多い。いろいろなタイプの統計調査でも、**母集団**と呼ばれている調査対象の全部を調べる代わりに、**標本**と呼ばれる一部分についての調査を行い、全体についての結論を帰納するという方式がよく用いられる。いわゆる**標本調査**と呼ばれている方法である。全体の素材が均質なものと考えられるならば、どのような一部分が調査対象に選ばれようとも結論にそう影響を与えることはない。だが多くの場合は、全体がかなり不均一なものであり、標本から得られる結論は、標本の選ばれ方によりいろいろと変動をする可能性をもっている。

いうまでもなく良い標本とは、母集団（全体）とよく似た性質をもっていることである。どのような選び方をした標本が、そんなうまい性質をもつことになるのか。この書では、母集団に対して信頼できる結論を帰納し得るよ

うな標本の、抽出法に関する理論の一端を紹介する。

標本抽出論に関する入門書や専門書は、これまでにいろいろ出版されている。大別して、理論の展開面を重視して記述されているものと、実際的な側面に重きを置いて書かれたものとに分けられる。理論と応用の両面をほどよくバランスさせたコ克蘭博士の名著(「サンプリングの理論と方法」と題された訳書も出ている)もあるが、やや大部に過ぎて初学者にはとっつきにくいきらいがある。邦書でも、理論的にある程度までしっかりと書かれたものとなると、やはり難解さが払拭しきれない感じである。

この書は、その書名が示すように、すべて例題から出発して諸概念の導入がなされるよう工夫を凝らしてある。例題は、实际的というよりはむしろ人工的なものである。抽出される標本のすべてのパターンをリストアップして、それぞれの抽出法の基本的な特徴をつかんでいただけるよう配慮してある。そのためには、母集団の構造を可能な限り単純化する必要があったからである。もちろん、より一般の場合で成り立っている公式を証明できる能力があれば、それに越したことはない。しかし、式の展開が理解できたということだけで満足していると、標本調査を実際に担当しなければならない場面になって、諸公式の使い勝手がよくわからないといった事態になるのが必定である。とにかく読者自らが、この書で提示された例題のようなものを作ることができるようになるまで、努力精進していただきたい。易しい、しかも意味のある問題が作れるためには、その理論のエッセンスを理解していることが前提となるからである。

ともあれ標本調査法は、実践あつての方法論である。理論を一通り習った後は、実際の応用に慣れるべくより多くの経験を積むことが肝要となろう。

昭和56年 盛夏

鈴木 義一郎

目 次

I. 標本調査の基本的な考え方

例 1.1	標本調査における標本	1
例 1.2	確率抽出法	2
例 1.3	標本調査の利点	4
例 1.4	標本調査の標本数	5
例 1.5	標本調査の手順	6
例 1.6	標本抽出作業の手続き	8
例 1.7	標本調査の誤差	9

II. データ記述のテクニック

例 2.1	パーセント	12
例 2.2	データの代表値	14
例 2.3	データの散布度	16
例 2.4	チェビシェフの不等式	18
例 2.5	多次元データの記述	19

III. 標本抽出の基本的な考え方

例 3.1	復元抽出と実験の独立性	22
例 3.2	非復元抽出と実験の従属性	24
例 3.3	平均値の確率分布(復元抽出の場合)	26
例 3.4	平均値の確率分布(非復元抽出の場合)	27
例 3.5	推定幅と的中率	28
例 3.6	くじ引きと乱数表	31

IV. 平均値の推定

例 4.1	標本平均の標本分布	33
例 4.2	ノーマル・チップス	37
例 4.3	正規分布	39
例 4.4	標本平均の統計的性質	42
例 4.5	標本平均の標本分布に対する正規近似	45
例 4.6	必要な標本数の見積り	47

V. 比率の推定

例 5.1	二項分布と超幾何分布	49
例 5.2	二項分布の確率と正規近似	52
例 5.3	比率の推定	55
例 5.4	必要な標本数の見積り	57

VI. 層別抽出法

例 6.1	層別抽出	59
例 6.2	全分散の分解	62
例 6.3	比例配分	63
例 6.4	最適配分	65

VII. 多段抽出法

例 7.1	多段抽出の有用性	68
例 7.2	集落の大きさが等しい場合の多段抽出(1)	69
例 7.3	集落の大きさが等しい場合の多段抽出(2)	71
例 7.4	確率比例抽出(1)	75
例 7.5	確率比例抽出(2)	78
例 7.6	副次抽出(1)	80
例 7.7	副次抽出(2)	82

VIII. いろいろな抽出法

例 8.1	集落抽出法	84
例 8.2	系統抽出法	88
例 8.3	層別集落抽出法	90
例 8.4	比推定	93

IX. 検 定 法

例 9.1	平均の検定	99
例 9.2	平均の差の検定	101
例 9.3	比率の検定	103
例 9.4	比率の差の検定	105
例 9.5	適合度検定	106
例 9.6	独立性の検定	108

解 答 111

参考書 127

付 表 (129～135)

I. 乱数表(129) II. 正規分布表(130) III. 二項分布表(132)

IV. t ー分布表(134) V. カイ 2 乗 (χ^2) ー分布表(135)

基本公式集 (136)

さくいん (140)

I

標本調査の基本的な考え方

人間には、自分自身も含めていくつかのパタンに分類しようとする性癖があり、日本人は特にその傾向が強い。だから、それぞれのパタンを代表する人達を集めて意見を求めれば、日本人全体の意向を推し測ることができる。全体(母集団)をよく代表する一部分(標本)をどのようにして選ぶか、これが問題である。

例 1.1 標本調査における標本

標本調査における標本はどのようにして選ばよいか。

◀ 解 答 ▶

調査対象集団(母集団と呼ばれる)のすべてを調査する全数調査に対して、その一部分(標本という)を抽出して調査を行うのが**標本調査**である。したがって、母集団の性質をある程度代表して説明できる標本を選ぶことが大切になる。

ところで、母集団をよく代表する標本を得るためには、母集団全体についての知識が必要である。実はそのような知識を得ようとするために標本調査を行うわけだから、母集団と何から何までそっくりの標本を得ようとしても無理である。調査目的にかなう観点がある程度制限したら、その意味での相似性を満足する標本を選び出すよう工夫することである。

まず考えられる方法は、全体の最も“典型的”と思われるような標本を抽出して調べるやり方で、これを**典型法**と呼んでいる。しかし典型的な標本を選ぶのが調査担当者であるとする、その人の知識や経験に基づく判断力に

依存する部分が大きい。また結果的にはかなり良いと思われる標本が選ばれたとしても、その標本に基づくある種の推定値が母集団の特性値とどのくらい食い違っているかを客観的に判定することができない。典型標本が母集団に対してどのような位置をしめているかが客観的に見当づけられないからである。

次に考えられる方法は、**割当法**と呼ばれているもので、これは母集団について事前にわかっている特性に関して、母集団と同じ比率となるように標本を構成していく方法で、一見良い標本を得る合理的方法のように思われる。しかし、割当法でいろいろな特性ごとに母集団の比率に似るよう標本を構成しようと意図しても、実際上はいろんな特性についてコントロールすることがむづかしく、また調査目的により直結した要因まで考えていこうとすると、結局母集団の特性がよくわかってないため、実際に近似させることが困難になってしまうのである。

以上、典型法と割当法について述べてきたが、この種の方法は調査担当者が意図的に標本を選ぶものであるから、**有意抽出法**と呼ばれている。一般に、この種の抽出法で標本を得ることは望ましいことではないが、客観性がそう問題にされないような場合には、ある程度の有用な情報が得られるかもしれない。

母集団の特性をよく反映し、しかも標本から得られる推定値の精度などまで客観的に評価できる方法が、いわゆる**確率抽出法**と総称されている各種の抽出法である。この種の抽出法の特徴としては

- (1) 標本誤差(標本抽出に伴う誤差)の程度が評価できる
- (2) 許容精度内の推論を得るための標本設計ができる

といった長所があげられる。この確率抽出法の導入によって、標本調査法は単なる全数調査の代用品というよりも、広範囲で適用可能な科学的方法を提供できるようになってきている。

例 1.2 確率抽出法

確率抽出法とは何か。

◀ 解 答 ▶

調査対象である標本を選びだす際に、何らかの確率を与えて抽出する方式の総称である。さて、ある確率抽出法を定めるとは

- (1) 可能な標本の組
- (2) 各標本が抽出される確率
- (3) 得られた標本から推定値を求める方式

の3つを定めることである。例えば、選択すべき標本の組として

$$\{S_1, S_2, \dots, S_M\}$$

が与えられたとしよう。これらは互いに異ならなければならないが、必ずしも可能な標本のすべてを含むものとは限らない。

次に各標本は

$$\{\pi_1, \pi_2, \dots, \pi_M\}$$

という抽出確率で選ばれるものとする。ここで

$$\pi_1 + \pi_2 + \dots + \pi_M = 1$$

という関係を満足していなければならない。特に

$$\pi_1 = \pi_2 = \dots = \pi_M = 1/M$$

の場合を**等確率抽出**という。最後に、 S_i という標本が得られたら、ある種の推定方式 $\hat{\theta}(S_i)$ によって、一通りに(母集団のパラメタ) θ の推定値が定められなければならない。 $\hat{\theta}$ の分布のことは**標本分布**と呼ばれるが、これが θ のまわりにより集中しているものほどよい。つまり良い推定方式とは

$$E\{\hat{\theta}\} = \sum_{i=1}^M \hat{\theta}(S_i) \pi_i = \theta \quad (\div \theta \text{でもよい})$$

という不偏性を満足し、さらに分散

$$V\{\hat{\theta}\} = \sum_{i=1}^M (\hat{\theta}(S_i) - \theta)^2 \pi_i$$

をできる限り小さくするものが望ましい。

確率抽出法の代表的なものとしては

- (1) 単純無作為抽出法

4 I 標本調査の基本的な考え方

- (2) 層別抽出法
- (3) 多段抽出法
- (4) 集落抽出法

といったものがあり、実際にはこの種の方法がいくつか組み合わされて用いられることが多い。これらの代表的なものについては、III章以降で詳説される。

例 1.3 標本調査の利点

標本調査の利点とは何か。

◀ 解 答 ▶

まず最も大きなメリットとしては、費用と時間が節減できることである。例えば全数の1%を抽出して調べようとすれば、全数調査の場合の費用のきっかり1%で済ませるわけにはいかないにしても、プラス・アルファ程度の費用を追加するだけで調査することができる。また、より少ないデータを処理するわけだから、調査結果を得るまでに要する時間が短くてすむというのも、かなり大きなメリットである。国勢調査は無論全数調査であるが、その1%集計結果というのが調査年度の翌年には報告される。しかし全数調査の結果となると、発表されるまでにはさらに数年遅れることになり、早く結果を利用したい人にとっては1%集計結果のほうが重宝される。

また標本調査による場合には、全数調査のときと比べて調査内容が拡大できるという点も特徴的である。例えば調査内容がかなり高度になるとより熟練した調査員が必要となるから、どうしても標本調査に頼る他はない。また特殊な装置の測定機構を利用する調査の場合でも、そのような装置の数には限りがあるから全数調査は事実上無理である。このようにみえてくると、全数調査は非常に限られた内容の調査しか行うことができず、逆にいうとたいがいの調査は標本調査に頼らざるを得ない。

さらに、標本調査による結果のほうが全数調査による場合より信頼できることも多い。標本抽出に伴う標本誤差が付加されるから、標本調査が全数調査より正確だというのは一見奇妙な感じがするかもしれないが、実はそんな

のである。調査には、標本誤差以外のいろいろなタイプの誤差——**非標本誤差**と総称されている——が入ってくるのが通例で、データを収集し、整理していく段階で絶えず監視していなければならない。このような管理機構を完璧なものにしようとする、データの数が多いことは逆にマイナスの効果をもたらすことになる。ほどほどの数のデータを得るというのは標本誤差を小さくする点からも必要であるが、無計画にデータ数を増やしても、とても使いものにならない悪質なデータだけを得るといった結果になりかねない。

以上、標本調査の利点をまとめてみると ①費用と時間の節減、②調査可能内容の拡大、③非標本誤差の減少 といった点が掲げられる。もちろん、標本誤差が伴うといった標本調査固有の欠点もあるわけであるから、標本の個数のあまりにも少なすぎるデータはやはり使いものにならない。しかし調査費用の観点などから抽出すべき標本数は限定されているので、可能な限り効率の高い(ばらつきの少ない)標本を得ることが肝心になる。良い標本を抽出するにはどうすべきかという問に答えるのが標本抽出論というわけで、この書はそのような理論の基本的な原理を明らかにしようという目的で書かれたものである。熟読玩味されて、よりよい標本設計を行っていただきたい。

例 1.4 標本調査の標本数

標本調査における標本数の決定は、どのような考えで行われるか。

◀ 解 答 ▶

標本調査による誤差(A)は、大きく分けると標本抽出に伴う標本誤差(B)と標本誤差以外の非標本誤差(C)から成っている。 B の大きさは標本数 n に関係(単純無作為抽出の場合には $1/\sqrt{n}$ に比例)して決まるもので、全数調査のときは0になる。一方 C の大きさは、全数調査、標本調査いずれの場合にも生じるもので、一般に標本数の増加につれて大きくなるものと考えられているが、どのように関係してくるかは明らかにされていない。

たいがいの調査では、標本数は調査費用の面から自動的に定まるケースが多いが、費用面で多少余裕があり、標本誤差の観点からみて妥当と思える標

本数を振り当てても、まだ若干の費用が残るような場合には、残りの費用を非標本誤差の大きさを見積るための再調査用にまわしたほうがむしろ適正な調査結果が得られるかもしれない。ともあれ標本数決定の際には

- (1) 要求された精度のデータを最も少ない費用で提供する
- (2) 与えられた費用で最も精度の高いデータを提供する
- (3) データを得る費用と、得られたデータを利用して行動したときのデータの誤差に基づく損失との和を最小にする

といった3つの原則に基づいて行わなければならない。

例 1.5 標本調査の手順

標本調査はどのような手順で進められるか。

◀ 解 答 ▶

一般的な標本調査の進め方としては

- (1) 調査の企画
- (2) 調査の準備
- (3) 調査の実施
- (4) 結果の集計
- (5) 結果の分析
- (6) 事後調査
- (7) 報告書の作成

といった順序で行われる。

(1) の段階では、まず調査の目的を明確に把握してから、調査対象集団(母集団)の概念をはっきりさせなければならない。また予算面等から勘案して、どの程度の規模の調査をどのような時期に行うのか、またどのような調査方法を用いるのが適切かを定める。さらに調査員の動員法であるとか、調査全体を通しての管理方法等についても、この企画の段階で策定しておかなければならない。

(2) の準備段階では、被調査者の選定、調査票の作成、そして調査員に

手渡す必要書類等の作成を行う。なおこの段階での作業、特に調査票の内容が調査結果を大きく左右するものだから、時間的余裕を持たせて慎重に準備することが大切である。

(3) の実施段階での最初は、調査員に対して調査を行う際の要領を説明することが必要で、これを**インストラクション**と呼んでいる。インストラクションは、調査担当者が直接調査員に指示を行うのが望ましい。また口頭による説明のほかに、調査要領等を簡潔に記した「調査員の手引き」のようなものを渡したほうがよい。なお調査現場では事前には予想できないようなことも起こり得るので、説明は可能な限り必要最小限に(ほぼ2時間くらいで)とどめておくのがよい。いよいよ調査員による現場作業——**実査**と呼ばれる——が進められ、調査票の回収が行われる。可能ならば一括回収よりも随時回収するのがよい。不明や調査不能理由をただしたり、必要なら再調査を命じることもできるからである。とにかく実査が行われている間は、調査員と調査担当者とは常に連絡できるようにしておき、問題が生じても中央からの統一のとれた指示に従って処理できるよう管理されなければならない。

次は(4)の集計であるが、最近ではたいてい電子計算機を用いるために、とりあえず結果をコード(付号)化する**コーディング**という作業から始められる。次いで計算機用のカードにパンチが行われ、いよいよ計算機処理が可能になる。まず質問項目ごとにどの答がどの程度あったかを見るための、**単純集計**が行われる。そして次は2つ(以上)の項目の間での関係をみるための、**クロス集計**が行われる。この種の集計は組合せの数が非常に多くなるので、性別、年齢別、職業別といった基本項目について、すべての質問とのクロス集計をとるのが通例である。この種の単純集計、クロス集計の結果を眺めて、調査結果がほぼ妥当なものであったかどうかを分析するのが(5)の段階である。特に調査不能者の集団にある種のかたよりがある場合には、調査結果全体にもゆがんだ結果を与えている心配がある。**フェイス・シート**(被調査者の氏名、住所、性別、年齢、職業、学歴、支持政党など調査項目とは直接関係のない質問群で、以前は調査票の最初に置かれていたのでこのような呼び名になっ

たが、最近では被調査者が嫌がる傾向があるというので後にまわすようになっている)の中で事前にわかっている項目について、調査不能のグループのかたよりの有無を調べてみることは、とりあえずやってみるべきである。結果のゆがみにかなり深刻なものが懸念される場合には、調査不能グループを主体とする吟味調査をしなければならないこともある。これが(6)の事後調査である。(7)の報告書の作成によって、標本調査の全過程が終了するわけであるが、報告書は分析結果の核心部分に言及し、一般利用者が利用しやすい形式に整理されていなければならない。特に調査目的からみて重点項目とみられる質問に関しては、そこで見込まれる誤差の程度についても明示しておき、将来同じような調査が行われる際に有用な補助情報を与えられるよう心しなければならない。しかし、なんとなく後味の悪い報告書を書かなければならないといった事態を避けるためにも、より良い標本調査ができるよう真剣に調査と取り組むことが大切である。

例 1.6 標本抽出作業の手続き

標本抽出の作業は具体的にどのような手続きで行われるか。

◀ 解 答 ▶

標本抽出の作業を具体的に行うには、調査対象である母集団についてもれなく書きあげられたリスト(選挙人名簿とか住民台帳のようなもの)を用意しなければならない。次はそのようなリストを、抽出操作の基本となる単位(これを抽出単位という)に分割する。例えば人間を相手の標本抽出の場合であれば、抽出単位が個人のこともあれば、世帯や国勢調査での調査区であったりする。また農業調査などによく用いられる地区抽出法の場合には、地図上で人為的に構成された矩形領域を抽出単位としてとられることもある。抽出単位に分割されたリストは、枠とかフレームと呼ばれている。とにかく抽出単位の枠は

- (1) 母集団の抽出単位をもれなく、しかも重複なく含んでいる
- (2) 枠で指定された抽出単位は、実際に弁別できしかも調査可能である

といった2条件を満たしていなければならない。わが国で5年ごとに行われている国勢調査は、このような枠を作るときのリストとして重要な役割を果たしている。

抽出単位の枠が定まったら、次はどのような段取りで調査用の標本を抽出するかである。現在では、例 1.2 で説明した確率抽出法の幾つかを適当に組み合わせて抽出が行われている。例えば、満 20 才以上の男女の意見を聞きたいといった場合には、枠として選挙人名簿を使うことができる。しかしこのような名簿の作成時期が調査時点よりもかなり前のことが多いために、ある程度(5～10%くらい)のずれは覚悟しなければならない。最初に抽出を行うのは調査対象の市区町村で、有権者の数に比例させて層別抽出を行う。この場合、国勢調査の結果などから利用できるのは人口規模しかないので、各地区の有権者数は人口数にほぼ比例するものと考えて、人口規模に比例させて抽出を行うのが常である。このように、1 段目に抽出する市区町村のようなものを**第 1 次抽出単位**と呼んでいる。さらに 2 段目の調査対象地区としては、投票区のように調査を実際に行うときの便宜を考えて広域にわたらないようなものを選ぶ。これを**第 2 次抽出単位**という。最後は、指定された投票区の中から選挙人名簿などを利用して、調査対象の個人を選ぶことになる。このように、3 段階に分けて最終標本を抽出する方法を**三段抽出**(一般には**多段抽出**)と呼んでいる。なお調査区から個人を選ぶ場合には、**単純無作為抽出法**と呼ばれる、いわばくじ引き方式のような方法で選ぶのがよいが、そのためには多くの乱数を使って抽出操作が面倒になるので、最初だけは乱数によって抽出し次は等間隔で抽出するという**系統抽出法**によって代用するケースが多い。

例 1.7 標本調査の誤差

標本調査の誤差としてどのようなものが考えられるか。

◀ 解 答 ▶

標本調査を進めていく段階を、(イ)計画段階、(ロ)観察段階、(ハ)整理段階の 3 つに分けて考えてみよう。