統計データバンクについて*

古瀬大六

1. 社会科学における統計データの重要性

社会科学もまた，経験科学であるという点においては，自然科学と異なるものではない。限られた経験に基いて論理的な矛盾を含まないさまざまな理論体系（モデル）が作られ，それが新たな経験のふるいに掛けられて科学が進歩していく。この点についてもまた，社会科学と自然科学とは基本的に同じである。

科学がその経験的事実を手に入れる手続きとしては，「観測」と「実験」との二つが可能である。自然現象についても「観測」は重要な方法であり，社会現象についても「実験」が絶対に不可能なわけではない。過去から現在まで，天文学や気象学はその大部分の経験を観測から得てきたし，最近の小集団理論はそのデータをすべて実験から得ている。

それにもかかわらず，これら二つの科学の間に方法上の大きな相違があるかのように考えられているのは何故であろうか。それは，決して，二つの領域の間に基本的・質的な違いがあるからというのではなく，相対的に言って，社会科学の分野ではコントロールされた実験が困難であり，自然科学の分野ではそれが比較的容易である，という事情に基くものである。同時にまた，自然現象はそれ自体が極めて大規模である場合でも，それを他の人間に危険を及ぼさないていどの小規模な実験に置きかえることができることが少くない。また，それを幾つかの要素現象に論理的に分解することによって，実験可能な知識の組合せとしてそれを再構成するということも，屢々行なわれる。

---

* 本研究は，昭和47年度文部省科学研究費総合研究（A）「日本統計の整合性と精度の研究」によるものであり，同研究関係者の御援助に謝意を表する．

他方，社会現象にあっては，実験を行なうこと自体が，一般の多くの人々に過大な危険と経済的負担とを負わせる結果となり，社会的に拒絶されることになる。また，小集団の実験データを積み重ねて大集団の理論を構成しようとする試みも，不成功に終りがちである。但し，このような特性が自然現象には全く存在しないというのではなく，相対的に言って，社会科学の対象とする現象には，実験を困難にする事情が伴いがちである，というだけのことである。

実験の困難さが，社会現象個有のものではないにしても，科学の進歩にとって，観測データの豊富に得られることが，欠くことのできない条件であることには，何の変りもない。社会科学が多数のグランド・セオリー（巨大理論）の対立状況にあり，経験による淘汰が極めて不十分であるのも，観測データの不足がその最大の原因となっているといってよいであろう。モデルをデータによってチェックする統計的理論と手法は，20世紀に入ってから急速に進歩をみせている。大量のデータについてこれらの統計学的計算を行なうことのできるコンピューターはすでに実用の段階に到達している。残された問題は，大量のデータを観測し，貯蔵し，社会科学者の要求に安価に応えることができるような具体的な対策を考え，実行に移すことだけである。そのような試みの幾つかはすでに実験的段階に入りつつある。

2. ミクロ・データにたいする要求

国民経済学の分野において，経験的データがモデルをチェックする役割を果すようになったのは，ケインズ・モデルによる所得分析理論が最初であった。周知のように，この種のモデルは，国民総生産・総投資・家計消費などのいわゆるマクロ諸量についてのモデルであること，従って，変数の数が少なく，モデルの形式も極めて簡単になること，をその大きな特徴としている。

変数の数を数個に抑えたことが，マクロモデルの方程式を簡単にし，従って，そのパラメターの値を統計的に推定することを可能にした。それがひい

ては，経済政策への実用に耐える理論を作り上げることを可能にしたのである。とはいうものの，これらのマクロ的変数の値を計測するためには，一人一人の所得，一つ一つの企業の投資を計測して集計しなければならない。マクロ数値そのものが客観的に存在するわけではないのだからそれを一つの測定値として測定することは不可能である。

また，政策的応用の面でも，今までのようなマクロ的政策が行きづまり，個々人の家計，個々の産業を対象に，いわゆる「きめのこまかい」政策を実行に移さないと，有効な政策効果を期待できないようになってきた。ここに，ミクロの理論と，それを支えるミクロの観測データにたいする要求が高まってきたのである。それにもかかわらず，従来のように，ミクロデータを集計して，そのマクロ的集計値だけを公表するというやり方は，貴重な情報の大部分を放棄することに外ならない。

事情は社会科学の他の諸分野においても同じであろう。人口調査，世論調査，意識調査，交通調査，等々のどれをとってみても，極端に少ないサンプル調査の場合を除いて，アグリゲーションによる情報喪失の量は極めて大きい。今後は，ますます大量のミクロデータが集積されるのであろうが，われわれはそれを最も有効に利用するためのさまざまな対策を予め考えておかなければならないであろう。

3. データの総合的利用

情報のロスをそれほど問題にしなかった理由としては，情報にたいする要求がマクロ変数についての情報であったこと，小標本の理論が未発達であったこと，をあげることができる。だからといって，現在の悉皆調査をすべて小標本調査ですまそうとすることは，社会諸科学の発展にとって望ましいこととはいえない。何故ならば，各種のミクロデータにたいする要求は今後ますます増加するであろうと思われるからである。とはいうものの，ミクロデータにたいする具体的な要求が発生したときに，その度ごとに大量のデータコレクションを高い経費をかけて行なうということは，容易なことではな

い。しかも，過去についてのミクロ・データを新たに集めることは，ほとんど不可能といってよいであろう。

ミクロデータにたいする要求の増大は，蒐集すべきデータの量を増大させる反面において，その増加率を減少させる要因をも生みだす。それは，今後ますますモデルがミクロ化していくと同時に，各変数の間の相互依存関係についてもますます細かい分析が必要となるであろうからである。従って，所得・消費・教育・意識・政治・地理的環境・文化的環境，等々のさまざまな分野の諸変数の相互の間の関係について，新たなモデルの作成とそのデータによる推定・検定が必要となってくるであろう。

換言すれば，従来の個別的要求に答えることだけを目的とする個別的データは，今後ますます，他の目的のために集められた個別的データとの関係という新たな視点に立って，利用される機会がふえてくるにちがいない。しかし，現在の個別的なミクロデータの内容そのままでは，これを他の系列のミクロデータと関連させることが不可能なことが少くない。居住形態と所得との関連を調べようとしても，住宅調査データと所得データとをミクロの段階で対応させるためには，この二つのデータ系列を同じ世帯主ごとに並べなおしてコレートしなければならない。現有のデータは，それに必要な情報を含んではいない。まして，アグリゲートされた住宅統計と，アグリゲートされた所得統計とからは，それらをどう操作してみても，研究者又は政策当局が要求するような住宅と所得との間の関数関係を求めることは，ほとんど不可能である。

異なるデータ系列からの情報を利用するには，各系列の調査対象を対応づけるに十分な情報を予め含めた形で調査しておかなければならない。この対応づけのための情報を「連結因子 (linkage or coupler)」と名づけるならば，すべてのミクロ調査に先立って，統一されたコードによるリンケージを個票に記録しておくことが必要であろう。

4. リンケージについて

リンケージの役割は，互いに異なる幾つかの一次元的なデータ系列を相互に関係づけて，それを多次元的空間のなかに再配列することである。あらゆる変量の間の相互依存関係を計測したいという要求をかなえようとするならば，リンケージの数もまた極めて多数にならざるをえない。従って，少数の最も普遍的なリンケージを予め用意するに留め，その他のリンケージについてはそれを基本的リンケージから間接的に作り出す，というのが現実的な対策というべきであろう。

そのための基本的リンケージとして誰が考えてみても当然に必要と思われるのは，時・人・場所の三つである。どのような計測データであっても，それが何時・どこで・誰について計測されたものであるかを明確にしておかないと，後の利用がむづかしくなるであろうことは，誰にも異論はないであろう。

「時」については，その精度がまず問題になる。国際標準時の分単位まで明記しておけば，正確なリンクを期待することができる。しかし，計測技術上それが困難な場合も少くないであろう。また，変量の性格上，あまり細かい測定をしてみても意味がないこともあるであろう。とはいえ，あまりブロードな単位を使うと，異系列間の時刻の対応がとれなくなる危険があるので，国際標準による時（アワー）単位を一つのスタンダードとし，必要により分又は秒単位まで降りる，というやり方が，基準形式として採用されるべきである。地方時による表示が望ましいならば，国際標準時からの簡単な換算によって，これを求めることができる。今後ますます国家相互間の交渉，全地球的な相互関係が大きくなるにつれて，計測時刻の世界的規模における標準化の必要は，大きくなるであろう。

変量によっては，所得・支出などのように，時刻ではなく時間（ピリオド）によって測る必要のあるものもある。これらについても，計測時間を標準化しておくことが望ましい。理想をいえば標準時から次の標準時までの一時間

を単位とすることにしたい。しかし，測定経費，其他の条件を考慮するならば，せいぜい，国際標準時の零時から24時までの1日間を採用するにとどめることにならざるをえないであろう。それでもなお短かすぎるというのであれば，週・月・四半期・年等の時間単位による計測もやむをえないかもしれない。だが，この場合に，計測の始めの時点を特定の国際標準時にそろえること，期間の長さをいつも一定に保つこと，長い方の期間単位が短い方のそれの整数倍であること，を配慮しないと，異系列間の時刻・時間の対応づけがむづかしくなる。これらの問題を解決することは，暦の改正とその国際的採用，にもつながる問題でもあり，決して容易ではない。それへの第一歩として，統計データのリンケージだけについてそのような国際的標準化を期待することは，それほど困難なことではないであろう。

第二の基本的リンケージは，人（又は事業所）である。調査対象が人間である場合には，異なったデータ系列を同じ人間について組み合わせたいという要求が生れてくる。それが可能になれば，個人の社会行動についての多くの理論の検証可能になるであろう。アメリカ合衆国では社会保障番号(Social Security Number)がそのような目的のために使われているし，北欧諸国では生年月日を含んで個人番号が，すべての国民にたいして，出生と同時に与えられる制度が，すでに実行に移されている。

この種の国民番号がすべての国民に与えられ，人にかんする統計データがすべてこの国民番号を附記して作成されるならば，既存の多くのデータを個人を中心として関連づけることができ，個人の行動様式とその歴史についての豊富な計量モデルを作り上げることができるであろう。アメリカにおいて，社会保障番号がそのまま納税書類の番号に使用され，国防省の兵役番号としても使われていること，周知の通りであり，最近の人口調査に際してもこの番号が併記された。ウイスコンシン大学によって行なわれた個人の投資行動に関する調査ＷＡＩＳ(Wisconsin Assets and Incomes Studies)は，この社会保障番号をリンケージとして，納税資料と社会保障資料とを結合することによって始めて可能となったのである。

個人に関する多くのデータを一か所に集めることは，その人のプライバシーに触れる可能性もあり，データの秘密保護に関する諸法規に違反する懼れもある。この点は極めて重要であるので，改めて後段において論ずることにする。

データが観測された「場所」については，従来は，主として行政及び郵政上の便宜によって設けられた「住所」によって表示されるのを通例としてきた。人間の手作業によって観測データを地図上に記す場合には，この「住所」による表示は，かなり高い効率をもっている。しかも，番地・街区・丁目・町・区・市という階層区分がはつきりしているので，アグリゲーションがやり易い。しかし，これをコンピューターに記憶させ，ディスプレイ又はプロッターによって自動的に図表化するということになると，「住所」システムは極めて厄介な代物に転化する。むしろ，ある一点を原点とする二次元座標を考え，全地域を等間隔の網の目で覆い，観測値がその縦何番・横何番の四角形の上に位置するかで表わした方が，はるかに処理が簡単になる。いわゆるメッシュ方式，あるいは，地域コード（ジオ・コーディング）方式がこれである。

道路計画・塵芥処理・交通調査・大気汚染データ其他のような，行政区画に制約されない地域データの表現形式としては，このメッシュ方式がその事後の処理にとって便利であり，この種のデータの必要は，今後ますます増大するであろう。

5. 統計データ・バンク

将来の統計調査が，社会科学者及び行政当局の多目的利用を目標として設計され，原始データのままで保存されたとしても，それを利用者にとって利用し易い形で提供するには，利用者の立場に立ってのさまざまのサービスが必要となる。

まず第一に生のデータに含まれている多くの誤記と未記入データとを補正することによって，利用者にとっての情報価値をできるだけ失わないように

しなければならない。原始記入がリダンダントなデータを含んでいるならば（例えば，収入の内訳と合計），その重複情報を利用して誤記をチェックし，或いは，記入洩れを補うことができる。単純な誤記又は転記誤りについても，それが常識的に取り得る値の上限・下限の外にあるときは，誤記である可能性が極めて高く，それを他のデータとの関連を辿ることによって，よりプロバブルな値に改めることができるであろう。

あるデータが完全に脱落している場合であっても，同じ調査対象について他の多くの項目が正しく記入されているならば，それらの諸項目の間に成り立っていると推定される関数関係を通じて，かなりの確度を以てその欠落値を補うことができるであろう。

これらの手続き（クリーニング）は極めて単調繰返し的な作業であるので，これをプログラム化することによって，コンピューターに自動処理させることが望ましい。修正又は補充された値にたいしては，特殊符号をつけることによって，これを生のデータから区別できるようにしておかなければならない。

第二に，提供を受ける研究者又は行政当局が，そのデータを直接調査した機関以外の人々であることを考えて，その調査の手続き，各項目の定義，クリーニングの方法，など，利用者にとって必要と思われる事項を十分明確に述べた解説（ドキュメンテーション）を予め準備しておかなければならない。日本人は，このような手続きを面倒臭がって真面目にやろうとしない性格をもっているので，特に強調しておく必要がある。この点における進歩を確実なものにする最も良い対策は，統計データバンク，或いは，統計データ・アーカイヴを独立の公的機関として設けることである。

第三に，データを保管する機関が，利用者の要求する再編集・統計処理を行なった結果を利用者に提供することが望ましい。これには，アーカイヴ自体が大型コンピューターを持ち，統計処理の専門家をかかえていることが前提となる。それは，現状では困難であるかもしれないが，将来はそのような形態にもっていくような努力をすべきであろう。

それが不可能であるならば，少なくとも，アーカイヴは，利用者の要求するフォーマットをもつた穿孔カードの形態で，データを提供できる施設をもつべきである。なぜならば，日本の大学・研究所・官庁は，そのほとんどすべてがなんらかのコンピューターを備えてはいるけれども，その機種は極めて雑多であり，磁気テープ，ディスク，紙テープの記録形式も互いに異なっていて，互換性を欠いているからである。お互いに障害なしにデータを交換できるのは，今のところ，穿孔カード以外にない。

穿孔カードとしてデータを提供するためには，アーカイヴ側では，簡単なコンピューターと，磁気テープ装置と，カード読取・穿孔装置とをもつ必要がある。これだけの設備があれば，カードを穿孔する際に，オリジナル・フォーマットを利用者の要求によって変更してカード化することは容易であるから，カード枚数が2万枚以下であれば，そのようなサービスを積極的に行なうことはそれほど困難なことではない。

コンピューター以前の研究者にとっては，既存統計資料への唯一の接近媒体は，「統計書」だけであった。コンピューター処理能力を持つ現代の研究者にとっては「統計書」は苦労の種でしかない。なぜならば，それは，パンチャーの手でカード化されて始めて，マシン・リーダブルなデータとなるからである。現在の穿孔カードは，コンピューターの機種の違いによる影響の最も少いマシン・リーダブルなデータ媒体である。しかしそれは，磁気テープからカードへの打ち出しの速度が毎分500から1,000枚ていどと，かなり遅いという悩みをもっている。磁気テープ又は磁気ディスクそのものが十分な互換性をもつような時代になれば，すべてのデータは磁気テープを媒体として利用者に提供されるようになり，入出力にともなう時間・経費・苦労が大幅に節約できるようになるであろう。

第四に，利用者の要求するデータの分類方式に合うようにそれを再編集するという作業は，場合によって極めて面倒な仕事であるので，その苦労を少なくするような工夫をしておかなければならない。そのためには，従来の統計書のように，いくつかの次元（キー）について，それを大項目・中項目・

小項目で分類・集計する，というやり方では，不十分である。データの管理機関は，自己の所蔵するデータについては，それを磁気ディスクに転記し，適当な（なるべく多くの）キーについての多重リスト構造を与えておくべきである。こうしておけは，利用者は，自分の特殊な要求に応じたデータ・ファイルを，マシン・リーダブルな形でアーカイヴから受取ることができるであろう。彼は，それを，自分の支配下にあるコンピューターにそのままインプットし，自分の管理下にあるプログラムをそれに適用して，解答を出すことができる。だが，この場合でも，データの管理者と利用者との間では，技術的打合せ，媒体の物理的運搬などにかなりの時間を割かなければならない。この労力を完全に省きたいのであれば，次にのべるような完全なコンピューター・ネットワークでセンターと利用者とを直接結びつけることにするより他に方法はないであろう。

そこで，第五の対策として，データ保有者と利用者とを通信回線で結ぶことを考えてみよう。この種のコンピューター・ネットワークには，大別して二種類の形態が可能である。その一つは，すべてのデータを一か所に集中管理するデータ・センターを設け，そこに超大型コンピューターを置き，全国に散在する多数の利用者の簡易端末装置と通信回線を経由してオンラインで結合する形態である。各端末からは，データの種類と加工プログラムの名称を送信し，すべての処理はセンターのコンピューターが引受け，結果だけを端末に送信する。加工プログラムは，典型的なものについては，センター側で予め用意しておいて，利用者はそのナンバーを指定するだけでよいことにしておけば，データ及びプログラムの送受信に要する無駄な時間と通信費用とを節約することができる。特殊プログラムについては，端末から対話形式でデバッグできるようにすることも，可能である。

第二の形態は，データの利用者が現在持っている計算施設をそのまま利用し，それらの間を通信回線で結ぶという形態である。各機関の手持のコンピューターの機種はさまざまであるから，相互の間で信号をやり取りし，必要があればコードを変換するための通信用小型コンピューターをそれぞれ一台

づつ用意しておかなければならない。また，各機関は相手のプログラムを自分のコンピューターの言語に変換するプログラムを用意しなければならない。これらの準備に若干の時間と費用とを必要とする。しかし，いったんネットワークが出来上ってしまえば，それは第一の形態と同じように使うこともできるし，各機関のコンピューターのうち手の空いているところを探してそこで処理する，というやり方も可能になってくる。

日本の大学・研究機関・行政官庁間には，そのようなネットワークは未だ作られていない。しかし，アメリカのＵＣＬＡ，スタンフォード，ランド，ハーバード，カーネギー，ＭＩＴ，イリノイ大，ケース工大，などの23の大学・研究所のコンピューターは，それぞれのインターフェス・メッセージ・プロセサー（ハネウェルＤＤＰ-516）を経由して，すでにＡＲＰＡ（アドバンスト・リサーチ・プロジェクツ・エージェンシー）ネットワークを形成し，活発な相互利用をすでに実行に移している。日本においても，同じ形態を作り上げることができない理由はないであろう。

6. 統計データのプライバシー保護について

わが国の統計法は，集められたデータがその統計の本来の目的以外の諸目的（例えば課税）のために使われることのないことを保証している。それ故，統計法に従って集計されたデータがデータ・バンクに集められ，社会科学研究或いは当初の目的以外の行政目的のために使用されるに当っては，法律上の制約を十分に考慮に入れておかなければならない。統計法によらない各種の調査であっても，調査当局と被調査者との間には，データの用途について，何らかの契約又は了解にもとづく制約が存在するのが普通である。

これらの法規又は契約上の使途制限の範囲をどのていどまで拡大できるかは，それぞれの具体的内容と情況とに依存するものであり，一般的な線を引くことは困難であろう。しかし，それを最も厳格に解釈すれば，最初に合意された目的を一歩も出てはならないという線が引かれると同時に，情報提供者の不利益になることが積極的に証明されない限りにおいて行政・研究上の

目的のために自由に利用することが許されるというもう一つの線を引くことができるであろう。現実の利用は，この二本の限界線の間で，適当な点に定まることになるであろう。

とすれば，統計データバンクの管理者としては，その保管するデータが，提供者の不利益にならないように十分な措置を講ずる最少限度の義務があることになる。ことにそれが個人についての情報である場合，いわゆるプライバシー保護の問題として慎重な対策が必要である。以下，データ提供者の保護と利用者の公開要求とを如何にして調和させることができるかについて，具体策を挙げて，考えてみよう。

第一の対策は，適当な暗号化によって，調査対象を特定化（アイデンティファイ）できないようにすることである。旧コードと新コードとの対照表（ディレクトリー）は，データ管理者が厳重に保管し，如何なる上級機関の要求があってもその提供を拒絶できる権限を与えておかなければならない。

第二の対策は，小規模のアグリゲーションを施した上で，データを利用者に提供する方法である。社会科学上の研究目的にたいしては，これで十分である場合が多い。経験上，三つのサンプルを一つに合計すれば，それから個別データを推計することは困難であるといわれている。また，集計値を使った場合に，それに各種の統計学的操作を加えて得られる結論がどれだけの影響を受けるかについては，今後の理論的・実証的研究を必要とするであろう。

(47. 12. 14)

文　献

1. Bisco, R. L., *Data Bases, Computers, and the Social Sciences*, Wiley, 1970.
2. Glaser, E. et al., The Design of a Federal Statistical Data Center, *The American Statistician*, Feb., 1967.