【フォーラム】（『統計学』第96号 2009年3月）

データ論についての覚え書き

― 内海庫一郎会員が提起したもの ―

森　博美*

序論

本誌創刊号の巻頭を飾るのは，故内海庫一郎会員による蜷川批判論文である。このことは，学会誌をこのような形で創刊した経済統計研究会の創始者諸氏の学問への意気込みを象徴するものとして興味深い。それは，単に学派の始祖に対する批判を掲載しているからではなく，論文の執筆者が，「蜷川統計学が昭和の初期に，その体系を構成しおわってから，すでに20年余りの歳月が経過しており，その間に統計的実践は多くの経験を積み上げて来ており，統計学とその直接的な関聯部門である哲学や経済学の研究も，大いに進歩した。…従って，現在の我々の課題は，この体系から出発しながらも，それを批判的に克服して発展させ，現段階の諸問題の解決に役立つように統計学を改造することでなければならない。」〔(2) 2頁〕として，社会統計学の新たな展開をめざしての論点提起を行っているからである。

この例に倣うなら，蜷川が「其の存在が社会的に規定せられたる集団」〔(1) 17頁〕とする「大量」に基づく演繹的理論体系として『基本問題』を世に問ってから既に75年が，また内海の論点提起からも45年余りの歳月が経過したことになる。その間に統計的実践は，新たに多くの経験を蓄積してきた。

蜷川の統計利用論は，わが国における二次利用の統計学のまさに嚆矢であった。政府統計の二次的利用が制度的にも現実味を帯びてきた現在，その本格的な構築が社会的にも要請されている。ところで，蜷川以降の本会会員による研究は，広範な二次利用者に対して具体的な利用の指針を提供しえたといえるであろうか。むしろ，統計的実践のその後の展開と本学会での研究活動の現状との間には，埋め難い溝が横たわっているようにさえ思える。

周知のように，戦後本会では，蜷川の後継者を中心に，学説史あるいはソ連における統計学論争等を踏まえて多岐にわたる立論が展開された。それは，統計学への標本理論や推測理論，また経済学へのマクロ計量モデルを中心とする数学的手法の導入に対して，統計学を社会科学における方法の学として確立することを目指したものであった。その中で蜷川理論についても，集団論，統計対象論，さらには統計学の学問的性格論として，蜷川に内在する現実許容的要素をそぎ落とし純化する方向での理論武装がはかられた。

この間展開された多岐にわたる論点とその評価を巡っては，すでに多くのサーベイ論文が著されている。そこで本稿では，統計の対象性を巡って提起されたいくつかの論点を参考に，統計情報，特に統計個票情報の情報特性にデータ構造面からアプローチすることで，社会統計学が今後取り組むべき課題について考えてみることにしたい。

* 法政大学経済学部
〒194-0298　町田市相原4342

1．蜷川の大量概念に対する内海の批判

蜷川は，統計方法の本質が，集団に関する数量的研究方法であると規定する。集団を単に個体の対立概念としてきた従来の集団規定に起因する理論的難点を克服するために蜷川は，集団を「大量」すなわち「其の存在が社会的に規定せられたる集団」と規定する。大量が社会的集団としての性格を持つことから，統計学は社会科学の領域における一つの研究方法論として位置づけられる。蜷川は，統計学の研究対象である統計方法を，何よりも集団，とりわけ大量という社会的にその存在が規定された集団に関する数量的研究方法であるとしている。

このような蜷川の大量概念に関して，内海は二つの興味深い論点指摘を行っている。その一は，大量の弁証法的性格に関するものであり，もう一つは統計対象＝個体説を巡る論点である。内海の問題提起を受けて統計対象を巡って展開された諸家の論説については，田中章義〔(4)〕に詳しい。なお，内海の蜷川批判は調査過程と解析過程の両面にわたるが，論点が拡散するのを回避するため，以下ではあえて調査過程に限定してその検討を行うことにする。

(1) 大量の弁証法的性格について

内海は，蜷川には「唯物論はあるが弁証法がない」〔(2) 2頁〕と批判し，社会的存在としての大量が事物論理として有する弁証法視点を大量概念に導入することで，蜷川理論に内在する二義性の排除を試みる。

このような立論の根拠となっているのは，「生成・発展・死滅的観点を蜷川統計学へ，まづ，その統計調査法に適用することを考」えた場合，「「存在たる集団」の単位標識は勿論のこと，時や場所の規定でさえも，それ自身，客観的事実に対応して，生成し，発展し，消滅するもの」〔(2) 4頁〕として捉える内海の認識である。なお内海は論文の補注で，大量を単に「社会的存在たる集団」と規定するだけではそれが現実に存在する「現象」であるという規定が十分反映されず，「大量」をむしろ「大量現象」と言い換えるべきであるとして自説を補強している〔(2) 12頁〕。

(2) 統計対象＝個体説について

内海は，「たしかに，統計調査は集団調査である。それは多数の個人の認識結果の総括，綜合である。」それでは「社会には個体は存在しないのであろうか？　それが単独に測られるという必要はないであろうか？」〔(3) 115頁〕，として蜷川が自明としていた統計対象＝集団説に対しても疑問を投げかける。内海にとって，統計対象としては，統計調査の時に調査票を配る単位としての個人，家計，企業といった社会的存在とその数量的規定があれば事足りるのである〔(3) 116頁〕。このような内海の立場からすれば，蜷川における「個体」は，大量の単なる構成要素として集団の中に埋没していることになる。

2．統計把握空間と統計

(1) 統計把握空間と個体情報

大量の構成要素である個人，世帯，企業等の個体は，一体どのような形で社会の中に存在しているのであろうか。それは，時間軸と横断面という二様の方向性を持って存在している。

象徴的にいえば，個体は時間軸上で様々なライフイベントを経験しつつdemographicに変貌する存在である。そこでは個体は生起（誕生）し，その時々に様々な部分集団を構成しつつ，時代という共通の場をくぐりながら，個々には時間の経過の中で着実に年齢（継続時間）を重ね，最終的には消滅（死亡）に至る。

他方，横断面では，各個体は，周囲の個体と様々な関係を取り結びつつ，同時点あるいは異時点での情報に基づき判断，行動する主

体として存在する。そこでは，個体の属性，行動それに意識といったものが一定時点における静態的存在，また時点間で存在形態を変える個体から構成される動態的集団として標識による統計的把握の対象となるだけでなく，個体間の関係そのものもまた個体と切り離し難い関連情報として存在している。各個体の属性や活動，さらには行動やそれを支える意識は，観測可能なあるいは観測不能な要因（変数）によって規定された形で存在する。横断面にあらわれる個体とは，まさにこのような情報特性を持つ個体なのである。そこでは各個体は，相互に孤立した無機的存在ではなく，時空を貫き相互に連関する社会的集団現象の諸側面を作り上げる主体として存在する。統計が究極的に把握すべき対象という意味で，このような個体から構成される社会的総体を筆者は「統計把握空間」と呼んでいる〔森(5) 16頁〕。このような視点から見れば，蜷川の「大量」とは，統計把握空間を一定時点で横断面方向に切り取り，その断面上に現れた静態的な集団性をいわば瞬間撮影によって捉えた姿であるとみなすことができる。

(2) 統計個票情報の情報特性

近代統計調査では，基本的に調査個票を用いて個体が保有する統計原単位情報の収集が行われる。そこでは，統計把握空間の中で様々な関係を相互に取り結び，判断し，行動する個体に帰属する情報要素が，調査票の調査事項として写し取られる。統計個票情報は，統計把握空間を構成する個体に関する統計原単位情報として，集計表に至る一連の過程の出発点とされてきた。

蜷川は，統計調査による大量の反映を，大量（大量観察）の四要素として，一定の存在の時と場所において標識について把握された単位に関する情報として捉える。しかしここで，統計個票情報の情報特性をデータ構造の面から特徴づければ，そこには蜷川とは異なる構図が浮かび上がる。すなわち，それ自体は非統計情報である調査単位の識別情報がいわばdata carrierとして，多次元ベクトルの形で個体に関する一連の変数情報（data body）を担っているというのがそれである。なお，ここでのdata body情報には，個体の属性や標識にあたる統計調査項目はもちろん，調査時点や把握の場所といった調査単位に関する時間や場所情報，さらには実査の場で調査員が独自に把握した調査区や個体関連情報も含まれる。data bodyが個体という調査単位そのものに関するcarrier情報によって担われるという統計個票情報の情報特性は，音声や画像さらには地域メッシュやGIS等が持つデータ構造と本質的に異なる。なお，この点の詳細は森〔(6)〕に譲る。

3．内海による問題提起の意味

(1) 統計＝個体説とアグリゲーション

より高次の集計量（マクロ）データの場合，ロビンソンの生態学的誤謬あるいはユール・シンプソンの逆説などとして知られるカテゴリーの統合（aggregation）に起因する新たなバイアスが発生しうる。また，マクロ（集計量）データとミクロ（個体）データによる回帰係数は一般に一致しない。この他にも，解析結果の意味づけを与える際に考慮すべき様々なバイアスが集計量に基づく分析には付き纏う。これらは，まさに統計の対象反映性の問題に他ならない。

蜷川以降の社会統計学では，集計に伴うアグリゲーションのレベルの問題は，部分集団の編成問題に解消され，社会科学の理論によって自ずと解決される問題として片付けられてきた。内海の統計＝個体説は，本人の意図はともかくとして，結果的には集計量を前提とした統計学が基本的に射程外としてきた個体差の検出とその除去，アグリゲーションに伴うバイアスの問題が統計学の取り組むべき課題として存在することを示したといえる。

⑵ 統計個票情報の情報特性とデータの潜在的拡張可能性

統計個票情報が与えるdata bodyが調査単位というcarrierによって担われる多次元ベクトルからなるというその情報特性は，個体識別情報をインターフェースとした異種の調査のdata bodyの接合利用の潜在的可能性を統計に付与する。

⑴ body変数の接合による次元の拡張

同一個体に関する異種のレコードを接合することで，data bodyの変数ベクトルの次元を拡張することができる。拡張された新たなdata bodyは，元のdata bodyがそれぞれ有する情報量の和以上の追加的な情報量を持つ。このようなデータの外延的拡張は，統計の作成面では，例えば静態データと動態データとの接合による新たなタイプの統計の作成を可能にする。また，利用面では，新たなデータセットは，それまでは行えなかった新たな変数の組合せによる多次元クロス集計あるいは重回帰分析を可能にする。

調査計画の不備あるいは調査負担の軽減等の理由で必要な変数が網羅できていない場合，集計量あるいは回帰分析結果は，系統要因の作用の不完全な統御あるいはそれを誤差として扱うことによるバイアスを持つ。この種のバイアスの除去可能性という意味で，data bodyの接合による変数の次元の拡張は，対象に対する認識の質の改善にも寄与することになる。

⑵ body変数情報の縦断的接合

「大量」が「生成，発展，消滅する」〔⑶ 250頁〕弁証法的性格を持つとする内海は，蜷川が「統計調査には一つの大量が対応」し，「存在を静止の立場でみるあやまった観念」に陥っていること，また蜷川が統計調査の基準的形態であるとする「悉皆大量観察法」が，「一連の歴史的な繋りにおいて存在するものの，時点的又は「短期的」な，一断面に限られた，全部的反映を与えるにすぎない」〔⑵ 5頁〕として次のように批判する。すなわち，「社会現象のある断面において，社会現象を構成している，運動している（生成発展，消滅している）単位の断面が…単位であり，断面はまさにこの単位の集団として理解される」。「悉皆大量観察は「大量」の全面的，多面的反映なのではない。それは一種の瞬間撮影に外ならない。言葉の本来の意味における全面性，悉皆性は与えられていないのである…。特定の大量の生涯の反映が与えられてこそ悉皆性の名にはずかしくない資料が得られるわけのものである」〔⑶ 251-2頁〕と。

それでは，内海はそれを反映しうる真の基準的調査としてどのような調査形態を想定していたのであろうか。いわゆる悉皆大量観察について内海は，「対象とその構成の変動が，緩慢で，かなりの期間にわたって余り変化しないときには，一度の断面図はかなり長期の構造の在り方を代表的に反映しうる。しかし，その変化が急激な場合には…かなり時幅のせまい断面図を作成し，静大量の時間的変化を系列化しなければならない」〔⑶ 252頁〕とする。悉皆大量観察の結果はあくまでも静大量であり動態集団ではないとしながらも，「短い間隔をおいてとった映画の瞬間写真をすばやく連続させると運動の印象が与えられるのと全く同じように，全系列運動の印象を与える。」とのフラスケンパーの所説を援用しつつ，結局は「多かれ少なかれ，このようにして対象の運動を模写する外はない」〔⑶ 253頁〕としている。ここで内海が想定しているのは，反復的な悉皆調査に他ならない。

異時点で反復的に実施される横断面調査の統計個票情報をリンクすることで，時間要素を内在させたdata body情報が得られる。個体を縦断的に追跡したいわゆるlongitudinal dataがそれである。個々の変数がこのような縦断的特性を持つパネルデータを用いれば，時点間の推移という時間要素を内包した明らかに静態量とは異なる動態的性格を持つ集団

構成が可能となる。また，時間に関して不変な個体間の差異を固定効果として抽出できるパネルデータを用いることにより，変数として観察不可能な要素も含め，その関与を統御した推計結果を得ることができる。

むすび

当時の内海の中心的関心事は，「統計学とその直接的な関聯部門である哲学や経済学」の最新の研究成果の統計学への導入〔(2) 2頁〕にあった。その意味で，大量の弁証法的性格あるいは個体性という内海の論点提起を，このように統計個票情報の情報特性という視点から捉えるのは必ずしも公正とはいえないかもしれない。しかし，大量という集計量に埋没していた動態的存在としての個体に注目し，それらの織り成す姿として統計を論じようとした内海の論点提起は，データ論としてはまさに時代先取り的なものであった。なぜなら，ミクロ解析手法については，生物学で1940年代半ばに提案されたばかりであり，一方，パネルデータについては，1960年代半ばにはじめて新たな統計のデータ形態として作成が開始されることになるからである。その後，欧米各国の政府統計機関の間で新たな統計の形態として普及するパネルデータは，集計量に基づく分析あるいは横断面，時系列データの解析からは得られなかった多くの新たな知見を提供してきた。

統計個票情報は，それが持つ特異な情報特性の故に，潜在的な接合可能性を持つ。リレーショナル・データが編成可能であるという意味で，統計個票情報のこのような情報特性をここでは，potential “relationality” と呼ぶことにする。統計個票情報のこのような情報特性からすれば，ミクロあるはパネルといったその後の統計作成あるいは統計解析法の歩みは，内海が蜷川批判の形で提起した，統計が反映する現実の弁証的性格と統計の個体性という二つの論点に対する現実の統計実践における一つの回答であったようにも思われる。

内海の問題提起を受けて，その後，本会の指導的理論家の間で様々な視角から論争が展開された。しかし，集計量に内在するバイアス，弁証法的調査のあるべき調査方式，あるいはデータ構造と情報損失といったその延長線上に存在すると思われる様々な論点については，残念ながら，その指摘もまたそれへの試行的取組みもこれまでは行われてこなかったように思われる。

統計調査や行政記録から得られる個体に関するデータは，いうまでもなく個体の全人格的描写情報を提供するものではない。また中には，その把握に技術的困難を伴う事項もあると考えられる。さらには，蜷川が信頼性，正確性として提起した反映すべき現実と反映結果との乖離も存在する。このような諸制約を持つ統計データから，われわれはどのようにして現実についての最大限のあるいはよりバイアスの少ない統計的認識を得ることができるのであろうか。また，時間情報を内在させた個体データあるいはそれらによる集団構成は，内海のいう社会現象の弁証法的性格あるいは瞬間撮影の連続映写による残像としての現実描写と一体どのように関係しているのであろうか。わらわれは，こういった問題を一つ一つ解きほぐしていく必要があるように思われる。

参考文献

［1］ 蜷川虎三（1932）『統計利用の基本問題』岩波書店
［2］ 内海庫一郎（1955）「弁証法と蜷川統計学についての一考察」『統計学』第1巻第1号
［3］ 内海庫一郎（1962）『科学方法論の一般規定から見た社会統計方法論の基本的諸問題』
［4］ 田中章義（1965）「統計対象にかんする諸家の見解について」『東京経済大学創立65周年記念論

文集』
［5］ 森　博美（2007）「我が国政府統計の展開と展望」『統計』1月号
［6］ 森　博美（2009）「統計個票情報の情報特性について」『経済志林』第76巻第4号