

あらためて「データ」について考える ——少し長めの解題

『日本労働研究雑誌』編集委員会

労働研究のみならず、実証科学と称する学術活動において「データ」が欠かせないことは、おおかたの論を俟たない。実際、本誌に掲載される論文の多くも、何らかの形でデータを利用して自らの立論に説得力を持たせようとしているし、学問それ自体が、まさにデータの取り扱いや解釈を巡って発展してきたといっても過言ではない。そして近年、データは学術研究のみならず、社会のさまざまな場面でますます利用されるようになってきている。現に社内外あるいはゼミナールのプレゼンテーションに役立てようと、データをあれこれ加工した経験をもっている読者も少なくないだろう。ほかに身近なところでは、新聞紙上やネットニュースでは失業率何%やら、最近では平均賃金がいくらだったといった統計数値が欠かせない。また、これまでデータと無縁だと思われていた法律分野も例外ではない。たとえば、「今後の労働契約法制の在り方に関する研究会」では、冒頭の第1回会議でいきなりA4判17ページにもわたるデータが資料（労働契約関係を取り巻く状況について）として提出されている¹⁾。極めつきは、政治問題と化した格差問題で、国会本会議場の代表質問でデータに関する議論が戦わされることも現実となった。

これほどデータについて喧しくなったのには、日本が名だたる統計大国といわれていることも関係しているだろう。たしかに、人口センサスの実施開始は遅れたものの、産業化以前の段階から貿易収支統計をきっちり揃えているのは世界広しといえども日本だけである。工業統計・商業統計・労働統計などの個別統計の数、調査継続年数も決して少なくない。本誌では10年ほど前の1995年1月号で「労働統計を読む」という特集を組み、当時の労働統計の現状と課題を整理したが、そこ

で取り上げられた統計は32種類に及んだ。そして、その特集の最後を飾る座談会において、小池和男氏がいみじくも述べたように、「日本の統計は、諸外国と比べると、個票の利用とパネルを除けば格段にいいというのが率直な印象」であろう。何かしらの議論をしようと思えば、どこからかそれなりのデータを引っ張ってこられるのである。

ところが、である。

実証研究が発展をとげる過程で、データを巡る真剣な論争を経験してきたがゆえに、そしてデータ利用が狭い学術サークルを離れて一般化してきたがゆえに、データにも高い信憑性が求められるようになってきた。データをとりまく環境は、10年前に比べると今日大きく変化したといってよい。そして、「格段にいい」と思われていた日本のデータでは本質的に対応できない要求がなされはじめ、それは奇しくも小池氏が留保した「個票の利用とパネル」に集約されはじめている。

まず、データの解釈を巡る議論の精緻化についてとりあげよう。たとえば、労働経済学の有名な命題に「ダグラス・有澤の法則」がある。有配偶世帯を考えたときに、主たる所得稼得者（多くの場合、夫）の所得が上昇すると、その他の世帯構成員（多くの場合、妻）の労働参加率が減少するという命題である。有澤廣巳氏自身が、1940～50年代の日本のデータを用いて、所得階層別の平均有業人員数をみたときに、所得階層が高くなればなるほど有業人員が少なくなる事実を指摘しており、集計データによれば、確かにダグラス・有澤の法則は成り立っていることが確かめられた。世帯所得と労働供給行動との関連という労働経済学にとっての基本命題が、ほかならぬ日本のデータによって発見されたことは、面目躍如といったところであろう。

しかし、集計データは個々の世帯で本当に上記のような関係が成り立っているかを担保しない。安定成長期に入り DINKS 世帯が脚光を浴びるようになると、自然にこのような疑問が提起された。このときは、集計データではなく「個票」を利用することで、同法則が個別の世帯でもおおむね成立することが確かめられたが、議論が終わったわけではなかった。高所得の夫と専業主婦の妻という組み合わせは、もともと専業主婦になりたい女性が高所得の男性をみつけて結婚している（またはその逆）だけであって、このような世帯では、仮に夫の所得が減少したとしても妻は専業主婦のままか、離婚して単身世帯に戻ってしまうだろうという反論がなされたのである。この反論が正しいとすれば、高所得の夫と専業主婦の妻という組み合わせの背後には、所得と労働供給との関係という経済学的なロジックは存在しないことになる。つまり、「ダグラス・有澤の法則」という見かけ上の関係の背後に、所得と労働供給という論理的因果関係を見つけようとする場合には、同一世帯で（しかもランダムに）所得の変動が起こったときに労働供給行動をどう変えるかを観察する必要がある。すなわち、パネルデータが必要となる。

データに関する議論の応酬がこの水準にまで達すると、日本に備えられていた格段によかったはずのデータは見る影もなくなり、一気にデータ後進国へ成り下がってしまった。現実には日本において実証研究は進まず、パネルデータを利用して「ダグラス・有澤の法則」が検証されたのは、ようやく 2000 年代に入ってからである²⁾。

もうひとつ、データ利用が一般化してきたことによって、データそのものの信憑性が問われるようになってきたのも、ここ 10 年の変化であろう。政府予算の策定に直結するためか近年まで公開されていなかった国民所得統計の推計過程が明らかにされるようになり、合計特殊出生率の推定にかかわるデータのサンプリングや計算方法まで一般紙でとりあげるようになった。データの利用者が増加することによって、利用者に対する明白な説明を公開する必要が強くなってきたと考えられる。

そもそも、データの信憑性をどのように確保するかという問題は、最近でこそ実験データの捏造問題が拍車をかけた感が否めないが、自然科学ではすでに Broad and Wade (1983) がいかに多くのデータ操作が行われているかを暴き、警鐘をならしていた³⁾。経済学でも同様の問題は焦眉の課題であり、Dewald, Thursby and Anderson (1986) は、*The Journal of Money, Credit and Banking* に論文が掲載された（または予定・審査中の）研究者にデータとプログラムの提供を求め、推定結果が再現できるか確かめた調査結果を報告している⁴⁾。そのなかで、とくに調査以前に論文が掲載されていた 62 名の研究者からは、回答が得られたのがわずか 42 名で、さらにデータとプログラムを提供したのは 22 名にとどまったことが明らかにされている。明確にデータ提供を拒否した 20 名のうち 14 名は、データは消失したと答えている。また、提供されたデータ・プログラムによって実証結果はおおむね再現できたとされているものの、著者たちは潜在的に再現可能なのが論文全体の 3 分の 1 程度であることに強い警告を発し、学術雑誌はデータとプログラムとともに論文の投稿を受け付けてしかるべきであると提言した。このような調査結果を受け、アメリカにおける経済学の学術雑誌の多くは、論文投稿時にデータおよびプログラムを提出するように義務づけられるようになった。査読者や編集者が実際に再現するかどうかは別として、潜在的に再現可能性を残すことで、データの信憑性を確保しようという狙いである。これに伴い、データをアーカイブに寄託し、公開するという方向も強固なものとなった。

ところが日本の対応はまたも遅れをとった。優れた統計の多くが官庁統計であったことも災いし、個票を公開する道は原則として閉ざされたまま、したがって日本を題材とした多くの実証研究は、外国の雑誌への投稿資格すら与えられていないのが実情である。国立公文書館こそ 1971 年に設立されているものの、収集対象となっているのは各統計の結果表にとどまっている。1998 年になっ

てようやく東京大学社会科学研究所にSSJ データアーカイブが設立され、民間のデータの収集保存が本格化した。官庁統計の個票を特別集計以外に利用保存する方法は、現在一橋大学経済研究所で試験的に行われているのみである。

本特集は、以上のようなデータを巡る環境の変化をまとめ、それに対する日本の対応を概括し、今後のデータ環境の整備と改善に資することを企図している。

まず、北村行伸氏の「パネルデータの意義とその活用——なぜパネルデータが必要になったのか」は、より厳密に統計学的な観点からパネルデータの有用性をまとめたものである。本解題では、エピソード的にデータに求められる情報量が変化したことをまとめたが、政策評価などの具体的な問題を取り扱うためには、統計学的な構造を把握することが望ましい。本論文は多少専門的ではあるが、労働供給の弾性値の計測や政策評価の枠組みを一望しており、よい導入となるだろう。

次の富岡淳氏の「労働経済学における主観的データの活用」は、近年の新しいデータ類型である主観的データについて、労働経済学での利用の実態を踏まえながらまとめたものである。ある人物にさまざまな形で将来予測や仮想現実について質問することで、回答者の期待を抽出する作業は今までも多く行われてきた。近年ではもう少し踏み込んだ「厚生」評価まで直接アンケート調査からとらえようとする努力もなされている。本論文では、こういった主観的データ固有のバイアスや誤差の問題、推定モデル利用時の注意点を整理している。もちろん、経済学における主観データの作成・利用は未熟で、経済学内部からも批判されることもあるほどである。著者自身も記すように「主観データの活用には……大きな可能性があると同時に、慎重な研究が求められ」、それゆえに、心理学など主観的データを扱ってきた諸兄の感想・知恵を待ちたい。

本多則恵氏の「インターネット調査・モニター調査の特質——モニター型インターネット調査を活用するための課題」は、字義通り、モニター型

インターネット調査が生むかもしれないバイアスについて考察したもので、同様な質問項目を調査法を替えて実施し結果を比較するというオーソドックスな研究の報告である。その結果、モニター型インターネット調査にはほかにはない「ある特徴」が観察される可能性が強いことを指摘している。本多氏は、サンプルサイズの点などから留保しながら、「調査の目的に応じて調査法を慎重に選択したうえで、可能な限りの誤差の縮減を図ること、どの調査法にも欠点があるという認識をもち、調査結果の利用の際にその欠点を織り込むことが重要」と結論している。研究者のみならず、現在モニター調査の利用者は少なくないと思われるが、実務家を含めて本論文を一読することをおすすめしたい。

佐藤朋彦氏と佐藤博樹氏の「データアーカイブの役割とSSJ データアーカイブの現状——実証研究における再現性を担保するために」は、東京大学社会科学研究所SSJ データアーカイブについて紹介するほか、米国のICPSR もとりあげ、社会科学におけるデータアーカイブの役割をまとめる。また、本雑誌に掲載された実証論文を取り上げ、当該マイクロデータを第三者が入手しその内容を検証できる条件があるかどうかの検討を行っている。その結果は本論文を参照していただきたいが、実証研究において再現性を担保することの重要性を考えさせる論考となっている。学問領域にかかわらず、実証研究に携わる方は是非ご一読願いたい。

坂本和靖氏「サンプル脱落に関する分析——消費生活に関するパネル調査」を用いた脱落の規定要因と推計バイアスの検証」は、わが国の代表的パネルデータで、もっとも頻繁に利用されているデータのひとつである家計経済研究所のパネルデータの脱落バイアスについて考察した論文である。米国のパネルデータでは脱落バイアスはあまり本質的な影響がないとわかっているが、当該データの場合、分析目的によっては深刻なバイアスをもたらす可能性が示唆されている。パネルデータを用いて分析する際に注意を要する典型的な例

といえよう。なお、本論文は特集と無関係に投稿され、通常の本誌査読過程を経由して掲載が決定された論文である。掲載決定後、本特集の主題と親和的なため責任編集者がまげて特集号での掲載をお願いした次第である。著者にあらためて感謝したい。

最後に配する吉川徹、永瀬伸子、樋口美雄、大竹文雄氏による座談会「『パネルデータ』を考える」は、現在同時進行中のパネル調査の担当者に集まっていたいただいた座談会である。パネル調査といえば坂本論文でも取り上げられた家計経済研究所のパネル調査が有名であるが、現在日本では複数のパネル調査が実行されつつある。その特徴を紹介していただいた。

以上、本特集は「データ」がおかれている環境が近年大きく変化しているとの認識のもと、その変化の内容・方向を見通すためにいくつかの論考・

座談会を取録した。これにより、多少なりとも日本の労働研究、ひいては社会科学研究の公正さが確立する一助になれば編集者冥利につきる。

- 1) もっとも、公開されている第1回議事録によれば、会議中にこの資料が触れられたのは、担当官が配布資料を説明するときただ1度きりである。
- 2) 武内真美子 (2004) 「女性就業のパネル分析——配偶者所得効果の再検証」『日本労働研究雑誌』No. 527, pp. 76-88. などがある。
- 3) Broad, William and Wade, Nicholas, *Betrayers of the Truth*, New York: Simon and Schuster, 1983. 邦訳に『背信の科学者たち』牧野賢治訳, 化学同人, 1988年がある。
- 4) William G. Dewald, Jerry G. Thursby and Richard G. Anderson (1986) "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." *The American Economic Review*, Vol. 76, No. 4., pp. 587-603.

責任編集 神林龍・小倉一哉・佐藤厚

(解題執筆：神林龍)