《資料》

海外データアーカイブの動向 4
―JSM 年次大会の報告から―

五十嵐 彰
髙橋 かおり

【要旨】 社会調査データは今後の社会の発展に寄与する公共財であり，広くデータが利活用される環境を整備する必要がある．本稿では 2020 年 8 月に開催された JSM 年次大会で報告された内容をもとに海外で行われている統計教育や実践の先進的な取り組みを紹介する．これらを踏まえ，今後の CSI 業務ならびに RUDA 運営に対しての方針と示唆を提案する．

キーワード：データアーカイブ，データ利用，データ共有

Ⅰ はじめに

社会調査部会は例年 IASSIST が開催する年次大会に参加し，社会科学に関する情報技術・データサービスに関する情報収集をすることを業務としていた．しかしながら必ずしも毎年同分野において大きな進歩があるとは限らず，また地域や手法に偏りが出てくる可能性を考慮し，今年は試験的に American Statistical Association が主催する Joint Statistical Meetings（JSM）に参加した．大会参加の目的は，海外，とりわけアメリカにおける調査技術や統計手法や，アーカイブのデータ活用に関する動向調査である．

今年は新型コロナウイルス感染拡大の影響でオンライン開催となり，出張は行わずに各自オンライン参加を行った．開催時期は 8 月 2～6 日であり、一部のセッションはリアルタイム配信，一部は事前に録画された音声と資料を提示する方式であった．リアルタイム配信は録画され一定期間公開されており，時差のある地域での参加も容易であった．

今回の年次大会のテーマは“Everyone Counts: Data for the Public Good”であり，公共財としてのデータの側面を強調したものだった．その名の通り，公的なデータの収集と処理，そして収集したデータを実践的に扱う方法に関する発表が多く見られた．またビッグデータや機械学習モデルなど，時流に沿った取り組みが多く紹介されていた．

IASSIST との差異とあげれば，アーカイブの管理側の取り組みやその運営に特化した発表はほぼ見当たらなかった．他方で統計手法やデータそのものの処理方法など，IASSIST では見られないが RUDA の業務に有用と思われる発表などもあり，従来とは異なる側面からの情報収集が可能になった．統計教育については IASSIST 同様充実していた．

本稿ではあまたある報告の中から，公共財としてのデータ活用，統計教育，web 調査，そして差分プライバシーの 4 つのトピックに絞って報告を紹介していく．

Ⅱ 公共財としてのデータ──商業目的と学術目的の間

JSM の今大会のテーマは「公共財のためのデータ」であり，セッションの中にも公共財としてのデータをどう活用するのか，データや統計を実践に生かす方法や事例についての報告が複数見られた．ここでは実践と教育という点から整理したい．

まず，実践についてはデータを他の専門職の業務や成果に生かす取り組みが紹介されたセッションがあった．「Date Science for the Public Good」のセッションでは，ジャーナリズムや法曹業界において，課題解決のためのデータ活用が実践例ともに報告された．ここでの調査課題は，親密な関係性における暴力や紛争での死者数推計，あるいは刑務所内での Covid-19 の感染状況など，今まさに問題となっている事例であった．これらの報告で興味深いのは，報告者やその実践への参加者の中で複数のデータを利用しつつ，いかにデータに書かれていないことを推測したり，データにあるバイアスへの注意を共有したりするということに重きが置かれていたことである．例えば，親密な関係性における暴力という問題について，個人単位，地域単位，週単位の 3 つのレベルで様々なデータを組み合わせることで問題の要因を探り，問題解決のための政策立案の素材としようとしていた．

いずれにしても，政府等によって公開されているデータの組み合わせで何がいえるのかということは，先鋭的な統計手法の発展とは別のやり方で統計やデータサイエンスの知見を深める方法である．これらの取り組みを可能にしているのは，データアーカイブの存在が大きいといえる．JSM の報告の中では「data analysis lifecycle」あるいは「data science lifecycle」という考え方がいくつかの報告の前提に提示された．これは，今日のデータアーカイブの議論の基礎になっている「data lifecycle」の考えの応用であろう（朝岡・髙橋, 2019）．

このような既存データを用いた問題解決のためにデータアーカイブとしてできることは，学術利用以外も視野に入れた構築や広報をすることであろう．現在の RUDA の規約では「教育・研究目的の二次利用」として，大学に所属する研究者の研究か，授業での教育が前提とされている．そして「商業目的での利用」は禁止されている．しかし，政策立案のための議論の材料としたり，何かの社会問題の解決のためにデータを利用したいということも今後考えられる．この「学術」と「商業」の間にある活用を考える際，今回の大会テーマである「公共財」としてデータを考えることはこの隙間を埋めることにつながるだろう．今日，EBPM（エビデンスに基づく政策立案）に注目が集まるが，EBPM でのデータアーカイブの活用は大学所属の研究者が利用することで「学術目的」としての利用と解釈されているだろう．しかし今後のデータアーカイブ全体のあり方としては，「学術目的」と「商業目的」の間にある「公共的な目的」をどう見るのかについて，留意しておくべきであろう．

なお，このような公的利用を考える際にヒントになりそうな取り組みとして，外部コンサルタントが調査倫理の相談を請け負っている事例の報告を取りあげたい．これは大学外のコンサルタントとして仕事を行う Kim Love が，データや調査における倫理的問題や困難についての相談をうけ，それにアドバイスをしているという報告であった．Kim は統計学での博士号を取得後大学に勤務を経て独立し，ASA の統計コンサルティングのセクションでも活動した経験がある．一方，学術業界ではないクライアントも対応している旨が自身のサイトに書かれており，まさに学術的なデータ利用とそれ以外の利用の両方をつなぐポジションにいる．そしてこのような仕事や調整にも専門家が存在しうるということは，公共的

な目的での調査データの扱いを促進していると推測できる．IASSIST におけるデータライブラリアンやデータアーキビストの実践例の報告からも顕著であったが，データを扱う際に研究者以外の専門職の充実を図ることもまた，今後のデータ活用において不可欠であろう．

Ⅲ　データサイエンス教育における考え方の伝達

次に教育の取り組みについて見ていこう．IASSIST においても統計教育の授業取り組みの共有に関するセッションが持たれていたが，JSM においても同様のセッションがいくつも組まれていた．ここでは高等教育（大学教育）において，統計手法の伝達よりもいかに自分でデータを探し，データ処理をし，可視化し，解釈するのかという点を強調するカリキュラムや実践について紹介したい．

通信環境や統計ソフトの発展，あるいはオープンデータ推進の流れに従い，だれでもアクセスできるデータが増えたため，授業内で教員が前処理をして提供するような整ったデータではないデータソースに学生が直接触れることが可能になった．そこで，統計学やデータサイエンスの専攻において，実験室的な状況ではなくより実践に近い形で展開される授業方法についての報告が複数なされた．それらの授業では，データを見る際の勘（intuition）やデータを扱う際の思考法を共有することが主眼に置かれていた．もちろん，ほかに基礎の統計教育の科目を履修していることが前提となるとはいえ，より実習的・実践的な取り組みを学会にて積極的に共有することは，興味深い．

例えば，学生が提出したデータとその解釈についてどのような間違いをより評価するのかという問いかけから，学生の評価の問題についての報告があった．そこでは，①正しい数値を導き出したもののその解釈を間違えた答案，②間違った数値を導き出したがその解釈は結果として正解と一致していた答案，③出題者が想定していなかった計算方法で導き出した数値から正解と一致する回答を導いた答案の 3 つがあったとき，この事例の報告者は③の回答を最も評価すべきだと考えていた．つまり，確かに学んだ手法を適切に使うことは大切だが，その思考方法や筋道の立て方を習得していれば，導き出しても構わない，という考えである．つまり論理的な結論を導けることは本来設定されていた仮説検定をするよりも重要な場合があるというのが報告者の主張であり，そのための教育を行っているのである．

このような実践の根拠として挙がっていたのは，アメリカ統計学会が出しているガイドライン GAISE（Guidelines for Assessment and Instruction in Statistics Education）レポートである．このガイドラインは，タイトルにもあるように，教授法だけではなく評価の方法も取り上げている．このガイドラインは日本の「統計教育連携ネットワーク」においても紹介されており，2005 年版は邦訳が公開されている [1]．ただし GAISE は 2016 年に改訂版が公開されており，ここで紹介する事例は 2016 年版の方針に基づく [2]。何を教えるかではなくいかに教えるかに焦点を当てた統計的思考の教育は，2016 年版に新たに追加された項目である。

JSM の教授法や授業実践に関する報告で共有されていたのは，論理的思考法としての統

計学をいかに教えるのか，そしてそれを学生が自主的に使えるように授業でどうすればいいのかということである．統計が技術や手法としてだけではなく，思考や哲学として扱われているともいえる．手法だけならば独学できるかもしれないが，どう考えるのかには研究者である教員やアシスタント，あるいは他の学生との議論が不可欠であろう．実際，データに関するディスカッションを取り入れた結果，それに積極的に参加した学生の方が成績は良かったという授業に関する報告もあった．

このように思考法としての統計データの活用として，CSI の業務との関連では，すでに実践している社会調査データ活用セミナーの取り組みがあげられよう．特に第 2・3 回目の応用編のセミナーにおいては，データの前処理の方法やその論理についての説明を重視した構成として行っている．単発の講座にはある程度限界があるとはいえ，独学であるとつまずきがちな考え方について共有する場は，オンラインでの自学自習ツールを補助するものとして考えられよう．あるいは，外部の研究者を招いてのフォーラムも，独学では難しい先端的な手法について，その考え方を中心に解説をしてもらっており，統計の考え方を共有しうる場となっている．

今後は CSI のセミナーやコンテンツを整理する中で各取り組み同士の連携を図り，調整や見直しも視野に入れた事業設計が必要であろう．今年はセミナーもフォーラムもオンラインで実施したが，次年度以降は対象者や難易度に応じて対面・オンラインを併用していくことも検討していきたい．

Ⅳ　Web 調査

CSI では従来（特にオンデマンド科目において）web 調査に関する問題点を指摘しており，無作為抽出調査を半ば絶対視する姿勢を維持してきた．しかしながら，web 調査は拡大の一途をたどっている．日本学術会議が「web 調査の有効な学術的活用を目指して」という報告書をまとめ，社会学評論では web 調査に関する特集が組まれている．国内外の査読論文でも，web 調査を用いた論文は増えつつある（e.g.，石田, 2016; Schachter, 2016）．RUDA にも web 調査会社を用いた調査の寄託が今後増加することが予想される．そのため，調査の特性，無作為抽出調査との比較，そして問題点の克服方法について整理しておく必要があるだろう．

1.　公的機関による web 調査の利用

Web 調査とは必ずしも完全に重複しないが，computer-assisted survey information collection（CASIC）の浸透についての報告が複数行われていた．CASIC とは従来の調査方法に対してコンピューターによる補助を施したものであり，例えばタブレット端末で調査票への回答を収集するものを指す．アメリカでは，consumer expenditure survey（日本でいう全国消費実態調査）の収集を実験的に CASIC で行っており，その成果が複数報告されていた．後述する国勢調査ともども，政府統計手法が今後大きく変わっていくと思われる．

アメリカで実験的に行われた consumer expenditure survey では，回答者が web の自記式，対面の他記式，電話の他記式の 3 つから回答方法を選択できるようになっている．web

の自記式は，家でオンラインにアクセスでき，週に 1,2 回はオンラインにアクセスし，英語話者であることが条件とされた．回答者の属性は，電話他記式を選ぶ回答者と web 自記式を選ぶ回答者の属性が似た傾向にあり，他方で対面他記式を選ぶ回答者はこれら 2 つの方式と比較し特徴的な傾向を持っていた．対面は高齢層，低学歴，低収入が多く，電話他記式・web 自記式は若年層，高学歴，高収入が多いようであった．エスニシティは白人が web 自記式を選択する傾向にあり，ヒスパニックは対面を選択する傾向にあった．これは言語の問題からくると考えられる．

web 自記式では，パソコンやスマホから一日の購入記録を入力できるようになっている．品目の大分類がアイコンで表示され，選択肢をたどるごとに細分化されていく（e.g.，アルコール，どのような品目だったか（ワイン，ビール，など），その値段，採ったのが朝食かどうかなど）．一日の総支出も閲覧できるようになっている．また consumer expenditure survey への参加謝礼を 5 ドル与えるか否かという追加実験を行っていたが，インセンティブは回答率に効果がないことが明らかにされた．

回答の負担感に関しては，自記式が最も低く，他記式だと負担感を感じるようであった．回答に費やす時間は回答モード間で大きな違いがなかったものの，web 自記式の回答者は所要時間を長いと思わない傾向にあるようであった．あくまで慣れであると思われるが，回答者によっては web 調査の方が負担が少ない調査モードであるといえる．

2. web 調査の誤差

Web 調査が必ずしも無作為抽出と相反するわけではない．例えば無作為に回答者を選択した上で，選ばれた回答者を web モニターとして登録してもらえば，無作為抽出調査といえるだろう．しかしながら現在の日本では，ほぼすべての web 調査会社では，潜在的な回答者が自発的に web モニターとして参加しており，調査会社はそのモニターのプールから回答者を募っている．こうした理由から現在の日本の web 調査はいわゆる無作為抽出調査とは異なっていると考えられる．

それでは，その違いが実質的な問題として扱われるべきかをどう決めればいいのか．JSM の報告の一つに，無作為抽出調査と web 調査の間の誤差を検出する方法を比較したものがあった．2 つの調査の変数の平均値と標準偏差を normalized root-mean-square error (NRMSE)と bias ratio (BR)とを用いて比較し，誤差の程度を測定するというものであった．

仮に無作為抽出調査と web 調査（非確率抽出）の間に大幅な誤差がある場合には，マッチングを用いて誤差の修正をすることが提唱されていた．無作為抽出集団間のマッチングの方法を比較検討している報告があり，そこでは K 近傍法，Hot deck，Quasi-randomization を比較していた．最終的には Quasi-randomization が最もよく誤差を解消することが示されていた．Quasi-randomization は大雑把にいえば，無作為抽出で収集されたサンプルを参照集団とし，web 調査で収集されたサンプルがサンプルとして選ばれる確率に重み付ける方法である．R の Survey というパッケージの svyglm，MASS というパッケージの StepAIC を使い実装可能ということであった．こうした方法を取り入れつつ，今後は RUDA においても web 調査のアーカイブを推進することが可能であると考えられる．

V 差分プライバシー

差分プライバシー（differential privacy）についての議論も盛んに行われていた．差分プライバシーとは，ランダムに収集したサンプルに対してさらにランダムネスを加える方法であり，公開された調査から個人が特定されないようにするものである．現在公開されている社会調査は匿名化がなされており，現段階でも十分に個人を特定できないようになっているともいえる．しかしながら，例えば Gary King が主導する Facebook データの公開や，国勢調査の学術利用，またはビッグデータに代表される個人の詳細なデータなど，オープンサイエンス化の流れの中で従来では公開されなかったデータが利用可能になってきており，プライバシーとのバランスを今まで以上に考慮する必要が出てきている（Dennis, et al., 2019）．差分プライバシーは，収集されたデータに対してランダムに発生させたノイズを加えることで，個人の特定を困難にさせる手法である．必ずしもすべての社会調査に対して応用される考え方ではなく，例えば European Social Survey といった既存の大規模社会調査に対する応用は現実的なものとして考えられていない（Oberski & Kreuter, 2020）．しかしながら一部の社会調査に応用され始めており，有名なところではアメリカの 2020 年の国勢調査に対して差分プライバシーが応用されることとなっている．

差分プライバシーの問題点は，測定誤差を増幅させることである．現にアメリカの過去の国勢調査のデータを用いて，差分プライバシーを応用することにより，非都市部に住むエスニックマイノリティの死亡率を従来の統計手法では正確に測定できなくなると報告されている（Santos-Lozada, Howard, & Verdery, 2020）．一方で差分プライバシーによりノイズが与えられたデータに対処する新たな手法も開発されており（Evans & King, 2020），こうした問題は乗り越える方向に向かうと予想できる．

実装可能性は一旦問わないとして，CSI にも差分プライバシーの考えが縁遠いわけではない．RUDA には地域データを多く扱う特徴があるが，こうした地域データは必ずしも完全にプライバシーが守られているわけではない．また立教大学の卒業生を対象にした調査などが一部演習で取られているが，プライバシーに配慮した上でこれらの調査を公開できるようになるかもしれない．また昨年の『社会と調査』には質的データアーカイブの構想が掲載されたが（髙橋, 2020），ここで最も重視・問題視されるのはプライバシーの問題である．差分プライバシーの応用により質的データアーカイブの構想を一歩前に進めることができるかもしれない．

VI まとめと今後への指針

本稿では JSM において紹介されていた取り組みをいくつか紹介した．従来参加していた IASSIST はデータを保存することに焦点を当てた発表が多いという印象であったが，JSM におけるデータを使う側の視点に立った発表にふれることで，データアーカイブの今後について検討することができるだろう．例えば，複数のデータを組み合わせた上で暴力事件の件数などを推測する方法などは，保存するデータの件数が多いほどより効果的な取り組みになると考えられる．その際には，本文中で指摘されているように，利用者の幅を広げるこ

とが重要といえる．現在 RUDA は利用を教育・研究目的に限定しているが，今後は利用の枠組みを広げることが重要となってくるだろう．仮に保存・配布するデータの件数に意義を見出すのであれば，web 調査といった非伝統的な調査データを収集・保存することも重要になってくるといえる．こうした調査は無作為抽出の観点から懐疑的な目で見られることが少なくないが，対応法を熟知・周知することによって web 調査の利点を活かすことも可能になってくるだろう．

統計教育という意味でも，大いに学ぶ面があったのではないかと思う．従来 CSI ではセミナーやコンサルティングにおいて統計手法に特化した取り組みを行ってきている．しかしながら実際には，手法と学術分野は切っても切り離せない関係にあり，手法のみを教えることは片手落ちと考えられてもおかしくないだろう．無論これは学部教育に深く関連している問題であるため CSI 単体の取り組みとしては限界があるものの，より学術分野に関連した，統計的・仮説検定的な考え方や議論の導出の仕方に重点を置いて教育しても良いかもしれない。

注 （ウェブサイトの確認日はいずれも 2020 年 12 月 25 日）

1) 「統計教育における評価と指導方法に関するガイドライン 大学レポート」（2005 年版日本語訳）https://jinse.jp/old/pdf/doc101.pdf

2) GAISE 2016 年版 https://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf

参考文献

朝岡誠・髙橋かおり，2019,「海外データアーカイブの動向 2――IASSIST 年次大会の報告から」『社会と統計』5: 33-40.

Dennis, S., Garrett, P., Yim, H., Hamm, J., Osth, A. F., Sreekumar, V., & Stone, B., 2019, “Privacy versus Open Science.” *Behavior Research Methods,* 51: 1839-1848.

Evans, G., & King, G., 2020, “Statistically Valid Inferences from Differentially Private Data Releases, with Application to the Facebook URLs Dataset.” Working paper.

石田淳，2016,「「日本人」の条件」『社会学評論』67: 182-200.

Oberski, D. L., & Kreuter, F., 2020, “Differential Privacy and Social Science: An Urgent Puzzle.” *Harvard Data Science Review* (in press).

Santos-Lozada, A. R., Howard, J. T., & Verdery, A. M., 2020, “How Differential Privacy Will Affect our Understanding of Health Disparities in the United States.” *Proceedings of the National Academy of Sciences*.

Schachter, A., 2016, “From “different” to “similar” an experimental approach to understanding assimilation.” *American Sociological Revie*w, 81: 981-1013.

髙橋かおり，2020,「質的データアーカイブ構想の現状と課題――数値化されていない調査データの保存と活用に向けて――」『社会と統計』6: 65-73.

Summary

The Trend in Foreign Data Archives 4:
From the Presentations at the JSM Annual Conference

Akira Igarashi
Kaori Takahashi

Social research data is a public good that contributes to the development of future societies, and thus we need to facilitate an environment in which people can use data appropriately. This paper reports on cutting-edge research and projects presented at the Joint Statistical Meetings held in August 2020. Based on these presentations, we suggest future directions for the Center for Statistics and Information (CSI) and Rikkyo University Data Archive (RUDA).

Key words: data archives, data usage, data sharing