

令和6年度 第2回図書館アドバイザー・レクチャー開催結果

- **日時** 令和7年2月13日(木) 14:00~16:00
- **会場** 県立図書館 学び⇄交流エリア
及びZoomによるオンライン開催 (県立川崎図書館 ミーティングルーム)
- **参加人数** 37名(県立図書館会場28名、川崎図書館会場9名)
- **アドバイザー紹介**
慶應義塾大学 文学部教授 岸田 和明 氏
慶應義塾大学文学部図書館・情報学専攻 教授。博士(図書館・情報学)(慶應義塾大学)。日本図書館情報学会会長。国立国会図書館・活動実績評価に関する有識者会議座長。駿河台大学文化情報学部助教授、教授等を経て現職。
専門分野は、情報検索、文書クラスタリング、テキストマイニング、図書館評価等。近年の論文として、「地域メッシュ統計を活用した公共図書館評価の試み：神奈川県を事例として」(共著 『日本図書館情報学会春季研究集会発表論文集』2022年度)、
「館外貸出に基づくマクロ指標による日本の公立図書館の状況把握」(『現代の図書館』59(3)通巻239(2021.9)等がある。

- **レクチャーテーマ**
図書館評価のためのデータ分析の技術

- **レクチャー概要**

今回は図書館評価の方法について概論をお話した。今回は、データ分析に焦点を絞り、来館者調査結果を分析した結果をお話する。

- 1. **図書館評価のポイント**

PDSまたはPDCAの中で、計画(Plan)と実施(do)が「主」であって、評価は「従」。評価を頑張りすぎると「本末転倒」になるので効率的にデータを収集・分析すべき。

複数の方法で評価を実施するというのが世界的な図書館評価の流れで、これをいかに効率的に行うかがポイント。近年、質的な方法の重要性が説かれ、質的・量的の併用が良いと言われている、組み合わせ方法については十分には確立していない。

- 2. **図書館評価のための測定技術**

歴史的な経緯を見ると、1980年代までは図書館学の教科書にあるサービスを着実に実現することが良いとされ、そこで使われていた図書館評価のデータが蔵書統計・貸出統計。もちろんこれらは今でも使われている。1980年代には、顧客ニーズから新商品を開発するマーケティングの考え方が導入されたが、新サービスの開発は難しく、読書

に対するニーズ調査などが主体だった。教科書には載っていないような新しいサービスを発案し図書館サービスの向上を図る取り組みは、もっと後の時代である。

1990年頃から営利企業の経営改善における顧客満足度という「アウトカム」概念が、図書館評価にも大々的に取り入れられるようになった。それまで図書館が満足度を問題にしていなかったわけではないものの、前面に出たのが1990年頃からになる。

2010年頃にサーチエンジンの高度化が始まり、この頃から「バリュー」や「インパクト」という概念が取り上げられるようになった。

図書館評価の測定は3本の柱で行われてきた。1つ目が貸出統計・業務統計、即ち日々の業務を遂行する中で発生するデータを集計した統計で、これに対しては数量的分析を行う。蔵書冊数、蔵書回転率、貸出密度などの指標が長く使われてきた。マクロ的な傾向を把握でき、コストが低く、統計的誤差が発生せず使いやすい利点があるが、分かるのは「量」のみという欠点がある。

それだけでは評価は終わらない、という問題意識から使われるようになったのが2つ目の、従来型のアウトカムである。データ収集には質問紙調査を用いる。来館者調査が代表的で、住民調査やモニター制度を使う自治体もある。数量的分析が中心になるが、業務統計に比べ質問項目を自由に設定できるという大きなメリットがある。満足度と期待度を合わせて聞くのが通例。様々な図書館サービスがある中で、期待度が高いものに対し満足度が高ければ良いし、逆に期待度が高いものに対し満足度が低いとアラートとなる。このように期待度と満足度のギャップを見るのが典型的な方法である。欠点は、調査者の方で固定した枠をはめてしまい「深く」は聞けないこと。その点を改善するために、3つ目の、バリューやインパクトを聞き取る、半構造化インタビューが使われるようになった。

インパクトとは、図書館サービスがどれだけ影響を与えることができたかということ。最近では、利用者個人だけではなく地域社会への影響も含まれ、地域をどれだけ活性化し、価値を高めたかという点でインパクトを捉えることがある。実際、様々な自治体で特色あるサービスを行い、地域活性化を図っている。

構造化インタビューは、質問紙調査と同様に、あらかじめ決めた項目を対面で聞き取るというもの。半構造化は、きっちり決めずにその場の流れで質問を変えていくもの。数量的分析に加え、質的分析という方法がとられ、インタビュー中に話された様々な事柄を、カテゴリーにまとめあげることにより事実やパターンを発見する。深く追求できるという利点に対して、標本サイズが小さく、一般化が難しいという欠点がある。その点、Plan、Doに対してバイアスがかからないよう注意する必要がある。

3. ISO16439：2014でのインパクトの測定方法

バリューやインパクトを測定する方法として、ISO16439という国際的な標準規格が2014年に制定され、図書館の評価に対するエビデンス（根拠）を3つ挙げている。1つ目はInferred evidence（推論されたエビデンス）。イベント出席、サービス利用のレベル等のアウトプットのデータや、図書館パフォーマンス指標、利用者満足度レベルを

通じてバリューやインパクトを推論する。直接的にはインパクトを測れないものの、数量的な尺度等から類推した間接的なものも組み込むべきという考え方である。2つ目は Solicited evidence (求められたエビデンス)。質問紙やインタビュー、フォーカスグループ・インタビューなどを通じて情報や意見を求めるという直接的な方法。3つ目は Observed evidence (観察されたエビデンス)。例えば定期的に館内を見回ることにより観察を行うような、構造化された観察やインフォーマルな観察、自分での記録、テストなど。

標準規格は以上のエビデンスを複数組み合わせることを奨励している。ただし、概念や定義づけが規定の中心で、具体的にどのように組み合わせるかの記載はない。

講師の個人的見解としては業務統計と質問紙の組み合わせが現実的。神奈川県立図書館は来館者調査を行っているので、標準的な方法を用いていると思う。インタビューの方法として、一般的には図書館協議会がある。学識経験者や教育経験者など図書館に関するステークホルダーを集め、報告事項、審議事項などと並び懇談事項として様々なことを話していただく。業務統計、来館者調査、図書館協議会の3点セットで図書館評価をやっていけばよいと、いろいろなところで申し上げている。

フォーカスグループ・インタビューの概要は資料のとおり。この手法を通じてバリュー・インパクトを測りたければ、図書館協議会の場でやるのが良いと思う。バイアスがかかる可能性はあるので十分な注意が必要だが、それを補うのが質問紙での来館者調査である。

4. データサイエンス (DS) と機械学習 (AI) の発達

文部科学省等が「数理・データサイエンス・AI 教育プログラム認定制度」(MDASH) を始めた。大学・高専卒業生、年間 50 万人全員に対し、DS (データサイエンス) と AI のリテラシーを身につけさせるという制度である。MDASH の認定をうけた科目に対して単位を取ると資格が与えられ、昨年 8 月時点で日本の大学の 50%以上が対応している。DS と AI に関する基礎的な知識を全卒業生に身に付けさせて、日本社会の発展に貢献してほしいという流れである。

生成 AI に関しては、2024 年前半に「Bing+GPT4」が登場した。Bing は Microsoft が提供している検索エンジンだが、そこに OpenAI の GPT4 が加わり驚いた。1990 年代の初め頃、Web を発明したティム・バーナーズ＝リーはページを探すのではなくページに書かれている情報をダイレクトに抽出したいと考えていた。そのための仕組みをセマンティック Web と言う。それが、2024 年前半に「通常の実用検索エンジン+GPT4」で実現した。自然言語の質問文に対して、ページから抽出した内容を要約し、それに典拠となる URL をつけることが「Bing+GPT4」がやったことであり、これはこれまで情報検索研究者がやりたかったことである。自分が大学生の頃にレファレンスの授業で「日本におけるマンホールの生産量は難問で答えられない」と教わったが、「Bing+GPT4」に尋ねてみると「正確な統計はありませんが…」という前置きで完璧な回答が出され、これからのレファレンスライブラリアンは大変だと思った。

Microsoft の GPT4 が先行したが、Google もそれに匹敵する Gemini という生成 AI を発表した。Google Colab を使うと自動的に Gemini を呼び出すことができ、勝手にプログラムを書いてくれるので、大学でプログラミングの課題は出せなくなったと同僚が話していた。それほど生成 AI は進化が著しい。

図書館界でも DS や AI については議論が行われている。図書館員に必要な DS のスキルと技術について、参考として『The Journal of Academic Librarianship』というアメリカの大学図書館向け専門誌に 2024 年後半に掲載された「Data competency for academic Librarians」という論文を紹介する。これは北米での大学図書館員に対する調査で回答者は 279 名。表 4 は自身の職（データライブラリアン、サブジェクトスペシャリスト等）と、各種データ関連業務（「図書館評価に関するプロジェクト」、蔵書構築、研究データ管理支援等）との関係を示している。また自己研鑽をしたいデータスキルを質問した図 3 のグラフが面白い。半数以上がデータサイエンスの中でも核となっているデータの可視化を選択している。次に人気なのが統計分析の知識、次に Python や JavaScript、MySQL などのプログラミングであった。Python は本当にすごいプログラミング言語で、生成 AI を可能にしたのも Python だと思う。

5. 神奈川県立図書館における評価データ（2023 年度）の分析例

神奈川県立図書館の来館者調査データをどのように分析できるかをお話する。データ可視化で出来る範囲や限界を把握して、PDCA の C（チェック）が図書館評価の根幹であることを実感してほしい。

来館者調査は、来館時に調査票を配布し、退館時に回収するのが普通のパターン。ISO11620 では 1 週間実施するよう規定されている。1 週間続けると、同じ利用者が何回も来館するため非回答率が上がるが、質問票を郵送するよりも高い回収率は得られると思う。ただし、来館者はある程度満足して来館しているので、満足度は高めに出る。非来館者を調査するには、自治体によってはモニター制度を使うこともあるようで、またお金をかければインターネット調査でも可能。

配付期間は、日本図書館協会の委託研究をした際に、平日と週末は利用者層が違うので両方必要だが、ある程度の標本サイズがあれば 3 日間で良いと『現代の図書館』に書いた記憶がある。バランスよく、バイアスがないように取るのが大切である。

データ可視化について。神奈川県立図書館の報告書中の「第 7 図 本日の来館目的」は複数回答で、棒グラフが用いられており、形式的には良い。性別は伝統的ではあるが、意味があるのかという気がする。「第 9 図 県立図書館を選んだ理由」も、形式的にはこれが一番良い。内容を見ると、場としての図書館が求められている感じを受ける。「第 11 図 本日の利用場所」も重要な質問で、これ以上はうまく視覚化できないと思う。「第 18 図 資料・サービスの満足度」は帯グラフを用いており、形式的には良い。項目別無回答が多いことが気になる。電子書籍等、特定の人しか利用しない項目が含まれるので当然だが、統計データとしては扱いづらい。4 段階回答を一覧するには帯グラフが妥当。加えて「満足」と「どちらかといえば満足」、「どちらかといえば不満」と「不

満」を圧縮して2区分にする方法もある。粗くなるので「良い悪い」を一概には言えないが、傾向の把握が楽である。

DS で取りざたされている視覚化がどれだけ使えるかを例示するために、Google Colab で Python を実行してみた。Python は比較的最近のプログラミング言語で、大変書きやすく多数のライブラリも存在する。生成AIは我々が普段使うCPUではなく、コンピュータグラフィックスなどで使うGPUを使わないと動作が遅い。Google Colabを使うと有料ではあるがGPUが使える。

Plan と Do で関わる部分で、効果的なクロス集計ができれば良い。性別はあまり関係がないと考えているが、年齢は関係がありそうなので、年齢別で掛け合わせて分析するとよいかもしれない。Excel の場合はピボットテーブルを使うのが普通だが、Python を使うと短いソースコードで、ピボットテーブルを自動で作成できる。

クロス集計はすごく大事。意味のある属性を選んでクロス集計できれば、何らかのパターンを読み取ることができるかもしれない。

「本日利用した場所」と「年齢」とのクロス集計の結果を可視化するための手法として、コレスポンデンス・アナリシス(対応分析)を実行して全体的な傾向を読み取ろうと試みた。これは二次元上にマッピングして、何らかのパターンを把握する方法として、最近よく使われている。原点に近いほど一般的な傾向、離れるほど特徴的で特異なパターン。項目間の距離が近いほど傾向が似ている。

次に、もう少し伝統的な箱ひげ図(Box plot)を紹介する。今の教育課程では中学生で学んでいる。テストの点数などのような、連続的な変数の様子を見るのに便利。図書館評価では選択肢が4段階程度のデータになるので帯グラフが分析に適しているが、お勧めのグラフとして紹介する。箱ひげの「ひげ」は、中学で学ぶのは最小値と最大値だが、最近の流行りはひげを短くし、外れ値を出して特異なデータを識別することが多い。

県立図書館の満足度について箱ひげ図を描いてみた。満足度はリッカート尺度であるから離散的で、4段階の平均はあまり良くないのだが出してみた。回答者はたいいてい良い方に回答するので面白い結果にはならないが、平均レベルで見ると項目ごとに少し差がある感じはする。箱ひげ図は全体的な分布を見るときに大変便利である。

次に、満足度を相関係数で分析した。以前は相関係数の数値が行列状に並んでいるだけだったが、最近はヒートマップを書くのが流行りである。0.0から1.0の間に相関係数が分布する。数値が高いほど赤く、低いほど青く、色を塗り分けて分かりやすくする可視化の手法でヒートマップを作成した。結果を見ると「開館日、開館時間」が他の項目と異なる。多分、他項目は満足でも「開館日、開館時間」は満足しないという人がいるのだろう。これについてはデータを詳細に調べればわかる。右下の「5.音楽・映像資料」から「11.生涯学習相談」のあたりが赤いのは無回答と関係がある。無回答が多く、回答者は皆、満足した傾向で答えたためだろう。質問が悪いのか、本当にそのような傾向なのかは分からないが、特異なので検討が必要だろう。このように異質なものが発見

されたら、確認して次回以降の調査に役立てるのがセオリーである。

次に、来館頻度ごとの満足度の箱ひげ図。来館頻度をどう使うかは考えどころで、「今日初めて」の人を外すことも考えられる。大学で授業評価アンケートを行う時は、出席率の悪い学生は、授業に対して分からずに回答しているという理由で外すこともある。来館頻度を尋ねるのはセオリーである。「今日初めて」を外してどれだけ意味があるかはわからないが、外したものと外さないものと両方分析するのが良いかと思う。

次に、主成分分析をかけた。「図書館の来館目的」と「県立図書館を選んだ理由」を合併しており、この回答に基づいた利用者のカテゴリー化が可能かどうか分析の目的である。大学図書館ではよくやる方法だが、2値変数で主成分分析をかけるのが良いか悪いかはよくわからない。結果では、丸で表された回答者がはっきり分類されている。似た傾向を持つ回答を重ねて聞いても仕方がないので、似ている項目は寄せて圧縮し、回答数や回答の負担を減らしたり、他の項目を追加して別のことを聞くなど検討すべき。

図中の丸の色は年齢が若い世代ほど薄くしてある。項目の距離が近いものが似ている。年齢で固まるかと期待したのだが、あまり関係はなく、むしろ「調査相談」「展示」「講座」など普通の利用の仕方と「視聴覚資料の利用」に二分されているようだ。

次に、満足度の高さに寄与する要因の特定ということで、ポアソン回帰分析を行った。満足度が高い人が何故満足したのかを分析したい場合、回帰分析を使うのが一般的である。「利用目的」と「県立図書館を選んだ理由」を一緒に入れているが、考慮なしに独立変数を多数投入しても、これらの間の内部相関のために結果の解釈が難しく、実は悪い例である。「満足：1」～「不満：4」で並べており、係数がマイナスだと満足度に効いていると解釈する。利き方の程度として、P 値が小さいほど効いていると見る。0.05 より小さくなると有意とするのが普通なので、今回の結果では1つだけ、「建物・雰囲気が入っている」人の満足度はほぼほぼ高い、ということが言える。回帰分析をかけると、このようなことが分かる。

最後に自由回答記述の分析を行った。BERTopic について紹介する。2017 年に提案された Transformer（深層学習の一種）を応用したアルゴリズムとして GPT と BERT があり、GPT は大規模言語モデルに基づく文書生成が可能で生成 AI で利用されるのに対し、BERT は図書館界では図書の自動分類などに応用されていてアメリカの医学図書館では盛んに研究されている。この BERT を利用した新しいクラスタリングの方法として BERTopic が 2022 年ごろ考案された。BERTopic を使うと大量の自由回答の内容を効率的に把握することが可能になる。

自由回答数は多くなかったが、結果を見ると BERTopic が自動的に解析し、肯定的評価と否定的評価に仕訳けている。回答数が多い時には要約もしてくれるので分析に便利である。現在、一部の研究者が、質的方法を生成 AI で自動化できないか、つまりインタビュー結果を文字起こしたものを利用して、人間がやるような質的な分析を機械ができないかを研究している。ある程度できるようになると思うので、半構造化イン

タビューを使わずに Google フォームのようなシステムで利用者の方に書いていただくと、そこから自動的にインパクト・バリューが読み取れるようになるかもしれない。

○意見交換と質疑応答

- Q1：フォーカスグループ・インタビューについて。インタビュワーの技術が求められると感じたが、この技術のレクチャーを受ける場はあるのか。場数を踏むしかないか。
- A1：学べる場はないと思う。大学院に行ってもフォーカスグループ・インタビューができる教員がいる可能性は少ない。類書は出ているので、本を読むしかない。インタビュワーの技術に影響されて結果が動いてしまう。
- Q2：標本サイズの考え方について。県立図書館のアンケートに適切な標本サイズはどのくらいか。例えば県民の数や、年間入館者数など、何を指標に割り出せばよいか。
- A2：ケースバイケース。サンプルがバイアスを含んでいるかどうかが大切で、偏っていたらサンプルを大きくしても意味がない。国の統計などでは無回答層に対してサンプリングして再度確認することもあるが、図書館ではできないので、標本が母集団に対しバイアスを含むかどうかを調査者がよく考えること。サンプルが 100 程度でも Plan Do のチェックに使えるのならそれで良い。適切な標本サイズを求める式はない。マーケティングで使われる式はあるが、母集団での分布のばらつきを仮定して入れないと出てこない。窓口担当者の感覚なども役に立つかもしれない。
- Q3：対応分析について、原点に近いほど一般的な傾向という表の見方で、19 歳以下が特徴的な位置にあるというご説明があったが、特徴的というのとサンプルが少ないというのはイコールなのか。40 代のデータと 19 歳以下のデータでは 19 歳以下のデータが少ないのではないかと思うが、そのサンプル数の違いが修正された状態でマッピングされているのかどうかお聞きしたい。
- A3：クロス集計をそのまま可視化しただけなので、19 歳以下のサンプルが少ない場合にはバイアスがかかっている可能性がある。サンプルを増やした時に 19 歳以下がもう少し 20 歳代の方に動くことは考えられる。ただ実際にデータを取って見ないと分からないので、標本サイズが小さいのはバイアスであると疑って解釈するしかない。
- Q4：県立図書館の満足度を分析した箱ひげ図のご説明で、面白くない理由として、リッカート・スケールで 4 段階だからというお話だったが、10 段階であれば良いか。
- A4：大学の授業評価もそうだが、皆さん優しくして良い評価になるので、箱ひげ図を書いてもあまり面白くない。ただ、やはり外れ値が出てきた時に、なぜ外れ値かということにはやはり役に立つ。10 段階にしても信頼性があるデータが取れるとは思わない。4 段階が良いと思うが、箱ひげ図としては面白い結果は出てこない。
- Q5：県立図書館の来館者アンケート集計データを視覚化する際のご説明で、プログラムで処理するとき Excel の上の部分が複雑すぎて苦勞したというお話があったが、Excel であればどのような形にすればデータ処理がしやすいか教えていただきたい。
- A5：上の方はセルの結合などで入れ子になっていた、Python で読み込むのに面倒くさ

かった。3行目のようだと全ての列にラベルが振れる。1列に必ず1つラベルがある形で質問項目があればもっと楽。セルの結合はしない方が良い。

Q6：男女を分けることに意味がないというお話が印象的だった。これまで統計的に「年配男性の利用が多い」など意味のある項目だと考えていた。分析をする上で男性と女性で回答に影響がなかったという意味か。

A6：Plan Do See なので、実際の図書館のサービスに役立つデータを取りたい。その時に男女を分けて何か役立つことがあるかという意味。むしろ例えば、仕事、育児などのライフスタイルで上手く分類できれば、開館曜日や時間に対する差が出たり、データベースや電子書籍は仕事をしている人の方が使うかもしれない、などサービスに連動する意味のある分類が、男女別よりもできるのではないか、という話。

Q7：満足度調査について、どの程度満足したかを聞くのではなく、期待度と満足度のギャップがどのくらいあるかを聞くというお話だった。これまで県立図書館では何を期待して来館したかという聞き方はしていなかったが、項目として入れた方が良いか。

A7：質問項目が増えると回収率や結果の信憑性に影響してしまうので避けたいところだが、期待度というか重要度を聞くとよくわかる。重要度が高いのに満足されていないところが分かると図書館サービスの改善につながる。質問項目数との兼ね合いか。

Q7：例えば、「何を目的に来ましたか」と「期待していることは何ですか」とでは、やはり違うと思った方が良いか？

A7：満足度と重要度は違う次元だと思う。やり方として、年度により大きな差はないと思うので、隔年で満足度と重要度を聞くことも考えられる。

以上