

データにもとづいた考え方とは(3) ー論理的に推論するー

東北学院大学 情報学部データサイエンス学科
教授 鈴木 努氏

第1回と第2回の記事では、事例の比較やことばのデータ分析の方法について紹介しました。最終回となる今回は、データにもとづいて論理的に考える方法やその際に気をつけることについて述べたいと思います。

論理的に考える方法としては、「三段論法」を思い浮かべる人が多いでしょう。「すべての人は死ぬ(大前提)」「ソクラテスは人である(小前提)」「よってソクラテスは死ぬ(結論)」という例が有名です。これは演繹と呼ばれる推論方法で、これに従えば論理的には常に正しく考えることができます。“論理的には”というのは、推論としては正しくても事実としては間違えることがあるからです。「すべての犬は骨が好きである」「ソクラテスは犬である」「よってソクラテスは骨が好きである」という推論は論理的には全く問題ありませんが、大前提や小前提の内容が事実としておかしいので、ナンセンスな結論になってしまっています。

データ分析との関係からいえば、演繹の推論は何らかの既知の法則から、データの予測を立てるときに用いられます。例えば、自動販売機での冷たい飲料と温かい飲料の販売量は季節によって変動することから、「気温が上がってきたので冷たい飲料を増やそう」「気温が下がってきたので温かい飲料を増やそう」などと販売の予測を立てることができます。

しかし、データにもとづいて考えるときに、より多く用いられるのは演繹よりも帰納と呼ばれる推論でしょう。これは観察された具体的事実から、一般的な法則を導こうとする推論方法です。そもそも、気温が上がると冷たい飲料が売れるという法則性がなぜ分かったかというと、過去のデータから売上と気温との相関が明らかになったからです。つまり、帰納によって導いた法則から演繹的に予測を立て、その予測に従ったデータが実際に確認されることで法則の確実性が増していくのです。この帰納と演繹の繰り返しによってデータにもとづいた考え方が洗練されていく流れは、ビジネスでも学術研究でも共通していると思います。

演繹と帰納に加えて、データにもとづく思考において重要なのが「アブダクション」という推論方法です。先の気温と飲料の例でいえば、飲料の売上を変化させる要因には気温のほかにも色々と考えられます。その中から

気温という要因に着目して気温のデータを集めるのは、それに先立って「気温が上がることで冷たい飲料の売上が上がるのではないか」という仮説があったはずで、このように、結果からその原因について仮説を立てるような推論がアブダクションです(米盛 2007)。

結果から、それを可能にする原因を推論しようとするとき、その論理は1つには定まりません。山の中の地層から魚の化石が発見された場合、「この地域はかつて山ではなく海だったのではないか」という仮説も可能ですが、「大昔、魚は山の中を歩いていたのではないか」という仮説も可能です。どちらの仮説がより適切かを判断するためには、その仮説を支持するような証拠、すなわちデータを集める必要があります。その地域の地質的特性をさらに調べて、そこが海だった証拠を見つけるか、陸上を歩くことのできた魚の化石を集めるかすればよいわけです。アブダクションは演繹のように、論理的な正しさが保証されているわけではないので、その正しさはデータによって裏づけられなければなりません。そして、観察された結果を説明するために、どのようなデータを集めればよいのかの手がかりを与えてくれるのも、アブダクションという推論の役割です。

仮説に従ってデータを集めた結果、仮説通りのデータが得られたとしても、気をつけなければならないことがあります。中学生を対象にした調査では、朝食を毎日食べている生徒の方が国語や数学の問題の正答率が高いそうです。このデータから「朝食を食べると成績が上がる」と結論することができるでしょうか。もちろん、朝食を食べることによって脳に栄養が補給され、試験に好影響を及ぼす可能性はあるかもしれませんが、しかし、中学生が朝食を食べるかどうかや、学業成績の良し悪しに影響を与える要因としては、家庭の養育環境や経済状況が関係しているとも考えられます。つまり、困難な家庭環境にある結果として、朝食を食べられない、試験で正解できないといった状況になっている可能性があります。

このように、2つの事象の間に関係があることがデータで示されている場合でも、それらが原因と結果として結びついているのではなく、共通の要因のもたらす結果として見せかけの関連が生じている場合があります。これを、相関はあっても因果関係はないという意味で疑似

相関と呼びます（毛塚 2022）。データで関連が見られた場合でも、その背景に共通の要因がないか気をつける必要があります。

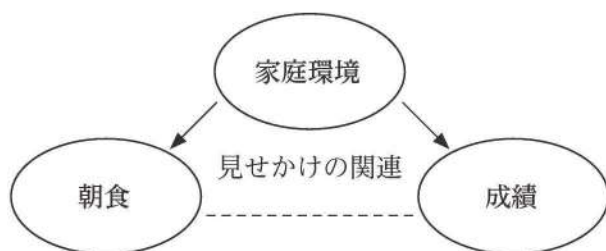


図1 擬似相関の例

同じような見せかけの関連は、時系列データでも起こる可能性があります。例えば月ごとの冷たい飲料の売上と水難事故の件数には関連が見られるかもしれませんが、両者に因果関係はないでしょう。もう、お気づきのよう、どちらも夏に増えるからです。同じように、時代の変化によって長期的に増えたり減ったりしている現象には因果関係がなくてもデータ上の関連が生じます。例えば、年ごとのタピオカ消費量の推移とロボット掃除機普及率の推移は、データ上では似たような動きをしているかもしれませんが、それは単に流行や普及の時期が重なっただけで、それ以上の関連はないでしょう。「タピオカ屋の店頭でロボット掃除機の宣伝をすれば売れるかも」という推論には十分な根拠はなさそうです。

タピオカ屋の店頭でロボット掃除機の宣伝をしようとするには、また別の点でも問題があります。それは集団における集計値の連関と個人の行動における連関を混同している点です。タピオカの流行やロボット掃除機の普及は、社会における多数の人々の消費行動を集計して得られるデータです。このような集計値の間に関連があったとしても、ある個人がタピオカとロボット掃除機をともに購入しているとは限りません。このような集団レベルの相関を個人レベルの連関と混同することで生じる誤った推論の例は、「ある食品の消費量が多い県ほど、ある病気にかかる人が少ない」「だからその食品を食べると病気の予防になる」といったような形で日常でもよく見られます。

私たちは、データにもとづいて論理的に思考しているつもりでも、自分の欲しい結果や仮説に合うデータだけに目を向けてしまう傾向があります。そのためデータの中にありもしないパターンを読み取ったり、そのパターンに合わないデータを無視したりしてしまいます。これを確認バイアスといいます（江崎 2020）。

近年、学術研究や医療のみならず、ビジネスや政策形成などさまざまな場面で、エビデンスが求められ、データにもとづいた意思決定が重要になってきています。データを用いて論理的に考える方法を知るとともに、その際に犯してしまいがちな間違った推論のしかたや、思考のバイアスの存在にも気をつける必要があります。

参考文献

- 米盛裕二, 2007, 『アブダクション 仮説と発見の論理』勁草書房。
- 毛塚和宏, 2022, 『社会科学のための統計学入門 実例からていねいに学ぶ』講談社。
- 江崎貴裕, 2020, 『分析者のためのデータ解釈学入門 データの本質をとらえる技術』ソシム。

〈プロフィール〉

福島県いわき市出身。東京都立大学で博士（社会学）取得後、東京工業大学、情報システム研究機構（国立情報学研究所）での博士研究員を経て、2013年より東北学院大学教養学部人間科学科教員。2023年、情報学部データサイエンス学科の開設とともに同学科教員として、主に社会調査やデータ分析に関する授業を担当。専門は人間関係などの「つながり」の構造を数理的に解析する社会ネットワーク分析。関心の中心は、人間・社会・文化・情報といった分野横断的な教育、研究。著書に『Rで学ぶデータサイエンス8 ネットワーク分析』（共立出版）、共著書に『テキスト計量の最前線—データ時代の社会知を拓く』（ひつじ書房）。

