統計的因果推論のイントロダクション

朝野熙彦
DS 研究会代表

【研究の背景】

DS 研究会はリサーチ・イノベーション委員会でデータサイエンスを研究する産学共同の研究組織である。このたび統計的因果推論の研究報告書を下記リンク先で公開した。本稿では因果推論で必要な予備知識を紹介して今回の報告書への橋渡しとしたい。

因果推論の方法論はローゼンバーグとルービンの傾向スコアに始まり、パールのグラフィカルモデリング、近年ではセミパラメトリックな SEM モデル、そして機械学習との連携へと進展してきた。

マーケティングでは、マーケティング活動の効果測定に常に関心をもってきた。近年では政府や自治体においても、EBPM（Evidence-Based Policy Making）といって行政施策の効果というエビデンスをもとに政策決定すべきだと強調されるようになってきた。因果推論の実務への導入は時代の要請であると思う。

【因果推論の予備知識】

ここでは報告書前半で必要になる予備知識を紹介しておこう。実は「相関と独立」を理解していれば大丈夫である。そんなことは分かっているという方にとって本稿は余計なお節介にすぎない。

まずは「無相関は無関係ならず」という警句から説明しよう。図1は気温を横座標 X とし、売上高を縦座標 Y にとった散布図である。

図 1 の X と Y は無相関なのだが、それにもかかわらず X さえ分かれば Y が分かるという明瞭な関係がある。外食ビジネスでは快適な気温のときに売り上げが伸びる傾向がある。暑すぎても寒すぎても人々の外出が減るからである。では X と Y のどちらが他の原因なのだろうか。この X→Y か Y→X かあるいは疑似相関なのかを相関だけで識別することはできない。なぜならピアソンの積率相関 r は $r_{XY} = r_{YX}$ なので因果の方向についての情報を与えてくれないからである。疑似相関のあぶりだしには報告書付録の偏相関が役に立つ。

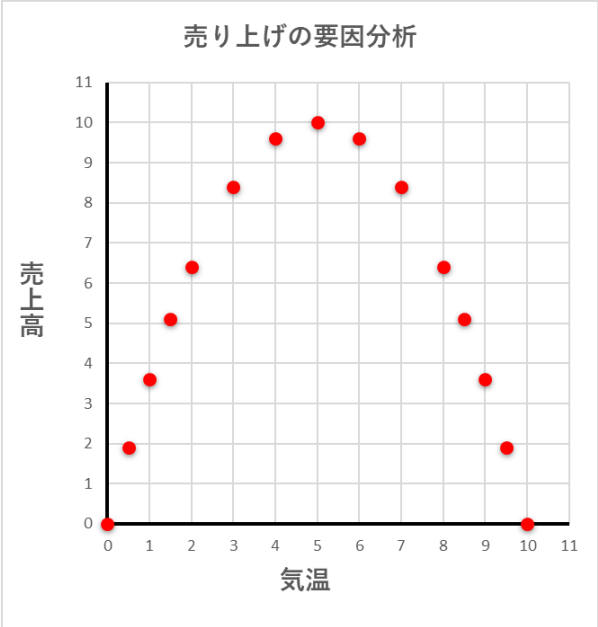

図 1　無相関だから無関係だと結論づけるのは誤り

無関係という表現は日常語なので、それを統計学でいえば独立という概念になる。2 つの確率変数 X,Y が独立であるとは次式がなりたつことをいう。

$$P_{XY}(xy) = P_X(x)P_Y(y) \qquad \cdots\cdots\cdots\cdots\cdots\text{①}$$

2 変数の同時確率がそれぞれの確率の積になっているときに X と Y は独立であるという。①式の右辺は同一の確率分布の積には限らない。その具体例を示したのが図 2 である。

図 2 の X は正規分布、Y はベータ分布にしたがう確率変数である。図2の X と Y は独立である。たとえば確率変数の値を X=0.8,Y=0.4 などと決めてやれば

X=0.8,Y=0.4 という座標点での同時確率 $P(X=0.8, Y=0.4)$ は周辺分布である $P_X(0.8)$ と $P_Y(0.4)$ の積で求まる。しかも、この 1 点だけではなく、XY 座標のすべての点で①が成り立っている。

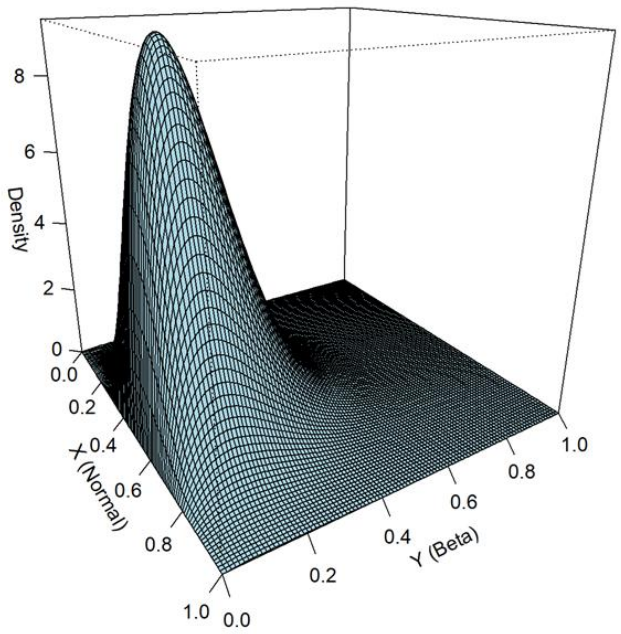

図 2　2つの確率変数が独立な場合の同時密度関数

ケーキを垂直にカットするイメージで図 2 を眺めてみよう。X の値をどこか 1 点に固定して分布をカットすれば断面はベータ分布になる。逆に Y の値を 1 点に固定して分布をカットすれば断面は正規分布になる。X と Y が独立であるとすれば②式が成り立つ。

$$P_{Y|X}\left(y|x\right)=P_Y\left(y\right),\qquad P_{X|Y}\left(x|y\right)=P_X\left(x\right)\qquad\cdots\cdots\cdots\cdots\cdots②$$

②式左辺の記法は条件付き確率を示し、今回の報告書の 2〜4 章に出てくる。このような連続変数の話では抽象的に思えるかもしれない。では離散的な変数で考えればどうだろうか。

表 1 は 100 サンプルのデータの 2 変数の同時分布を示した例である。X のカテゴリー数を 3 つ, Y のカテゴリー数を 4 つにした。表 1 では変数 X,Y が独立になっている。たとえば最初の（X1,Y1）のセルを見ると、100 分の 15 だから確率でいえば 0.15 である。この 0.15 は X1 である確率 0.5 と Y1 である確率 0.3 を掛けた結果に一致する。そして表 1 のすべてのセルにおいて①式が成り立っているので X と Y は独立だといえる。

表 1 は2つの質問のクロス集計表に他ならない。そしてここで述べた独立性を確認する計算はカイ二乗値を用いた「独立性の検定」の途中計算と等しい。表1でカイ二乗値を計算すると $\chi^2=0$ で 2 変数が独立であることが確認できる。

表 1　独立な離散変数

	Y1	Y2	Y3	Y4	合計
X1	15	20	10	5	50
X2	12	16	8	4	40
X3	3	4	2	1	10
合計	30	40	20	10	100

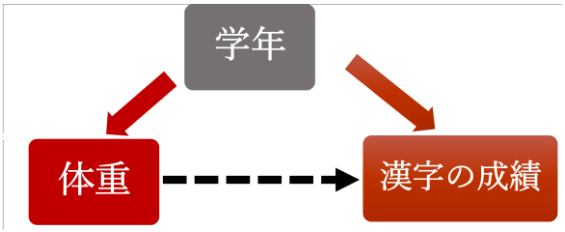

図 3　体が大きい子ほど漢字ができる

次に因果グラフと独立性の確認である。図3は、「小学生は体が大きい子ほど漢字ができる」という因果グラフを表している。体重と漢字の成績には通常正の相関がある。因果推論では次のようにこの問題に答える。学年を条件として固定すれば体重と漢字の成績は、おそらくは独立だろう。もし独立だったとしたら図3の破線の矢印は削除していいはずだ。図3の例で条件付き独立が因果推論でどう使われるのかが納得できるだろうか。

【研究報告書の意図】

今回の報告書では様々な因果推論を実行する Python や R のコードを記載している。しかし、本報告書はコードをコピペして実行できればそれでおしまい！というような「写経の書」を意図したものではない。因果推論の論理を考察し、応用の前提条件を明らかにすることをめざして研究メンバーは努力している。

なお今回の研究報告書には 2023 年の研究会で扱ったマハラノビスの汎距離が因果推論の一方法として登場する。そもそもマハラノビスは考古学の研究のために汎距離を考案したのだった。まさか汎距離が因果推論に使われるとはマハラノビスは予想もしなかったに違いない。このように異なる応用分野の舞台裏で共通の理論が働いていることは珍しくない。たとえばスパース回帰分析や KL 情報量はデータサイエンスの随所に出てくる。DS 研究会の過去 2 回の報告書にはデータサイエンスのさまざまな理論が登場するのであわせてご覧いただきたい。

今回の報告書が企業のマーケティング戦略や行政の EBPM に貢献できるなら幸いである。

（東京都立大学元教授、リサーチ・イノベーション委員会委員）