コラム：相関と因果

相関係数というのは, 2 変量データ (x, y) の関係の中で, 特に線形関係 $y = ax + b$ を想定したとき, あくまでもその線形性の度合いを測る指標となっています. ですから, 2 変量データ (x, y) が一直線上に並ぶときには, 相関係数の絶対値は 1 となり, それ以外では 0 から 1 の値を取るのです.

一見 2 変量データ (x, y) の相関係数が高いと, 2 変量に関係性があるように思いがちですが, 別の変数 z が両者に関係することによって生じた疑似相関の可能性もあります. この変量 z のことを交絡変数といいます. 有名になった例としては, 人口当たりのノーベル賞の受賞数とチョコレートの消費量の関係というのがあります. この 2 変量の相関係数は 0.79 もあり, 一見するとチョコレートを多く消費するとノーベル賞を取りやすいかのように見えます. ここで 2 変量に関連する変数として, GDP や大学進学率などが考えられますが, z を GDP として計算したところ, 2 つの相関係数は 0.32 となり, さして関係性がないものとなりました. これを疑似相関といい, この時の変数 z を交絡変数といいます.

特に相関関係と因果関係は意味するところが違いますので, 気を付けましょう. 因果関係の定義としては, 18 世紀の哲学者であるヒュームによる以下の 3 つの条件：①原因と結果が空間的・時間的に近接していること, ②原因が結果よりも時間的に先行しており, 継続して結果が起こること, ③第三の要因が同じである場合に同じ原因から必ず同じ結果が生じること, が有名です. この 3 条件を踏まえると, 自然科学の中で実験が可能な分野では, 独立変数の操作で従属変数が変化することを実験する独立変数の操作性や, 独立変数と従属変数の時間的順序性などが重要となります.

このような実験が出来ない場合, 例えば, 塾に通った時と通わなかった時でのあなたの成績の伸び具合の差を知りたいと思った場合, あなたがもし塾に通うならば通わなかった場合を考えることが出来ず, もし塾に通わなければ通った場合を考えることが出来ないといった二律背反な状態に対して, 事実と反する結果を想定する必要が出てきます. これを, 統計的因果推論における潜在的結果変数を想定した反事実モデルといいます.