2022年6月17日　統計数理研究所　オープンハウス

集計データの統計解析（エコロジカルインファレンス）

岩崎　学　　大学統計教員育成センター　特任教授

1．オープンデータとエコロジカルインファレンス

昨今の統計・データサイエンスの広がりと社会からの期待には目を見張るものがある．社会に提供されるデータの量も種類も爆発的に増加しつつあり，それを分析する職種としての統計家あるいはデータサイエンティストが脚光を浴び，その数も爆発的に（かどうかはわからないが）増加しつつあるのが現状である．昨今のデータ事情の特徴としてオープンデータの提供がある．これまで官公庁が独占していた各種の調査データが，データは公共物との観点からオープンデータという形で広く公開され，それを政策決定やビジネスなどに広く用いる可能性が出てきている．そしてそれが，データを扱う専門家としてのデータサイエンティストの増加に拍車をかけている．

オープンデータの特徴は，それが集計データであることである．実際，政府の提供する各種調査データは都道府県別や市区町村別に集計されたものがほとんどである．また少し違う形でのメッシュデータも同様である．もちろんそれは個人情報保護の観点からも重要であるが，分析段階ではそれが集計データであることを忘れてはならない．集計データの解析法としてエコロジカルインファレンス (ecological inference) があり，それは今日さらに重要性を増しつつある，というのが筆者の見解である．しかし，エコロジカルインファレンスに関しては，本邦ではあまり書かれたものを見かけない．岩崎 (2019) の第3章あるいは岩崎 (2022) はその数少ないうちのものとなっている．一方海外に目を向けると，単行本としては Achen and Shively (1995)，Bossarte (2008)，King (1997)，King., Rosen and Tanner (2004) などがある．

ここでは，簡単な数値例を基に，エコロジカルインファレンスの特徴を考察する．

2．数値例2つ

例1（大学の学科における性別と英語力）

これは，筆者が過去に実際に遭遇したデータに若干手を加えたものである．表1は，文系理系の両方を有するある大学の9つの学科における女子学生の比率 (x) と，大学で実施された英語検定試験で合格水準に届かず不合格となった学生の各学科での比率 (y) である．図1には，表1のデータを2次元プロットし，回帰直線を描き入れている．

例2（ワクチン接種率と重症化率）

数値は例1と全く同じとした架空データで，対象をある疾病のワクチン接種率と病気の重症度としたものである．表2は，9つの異なる地域におけるワクチン接種率 (x) と病気の重症化率 (y) であるとし，図2がその2次元プロットである．

表1．女子比率と不合格率

学科	女子比率	不合格率
1	18	38
2	29	30
3	30	37
4	33	28
5	35	30
6	70	32
7	72	27
8	79	23
9	77	18

表2．ワクチン接種率と重症化率

地域	接種率	重症化率
1	18	38
2	29	30
3	30	37
4	33	28
5	35	30
6	70	32
7	72	27
8	79	23
9	77	18

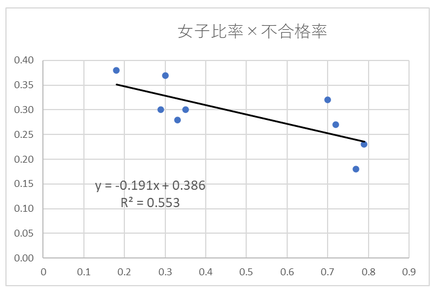

図1．女子比率と不合格率のグラフと回帰直線

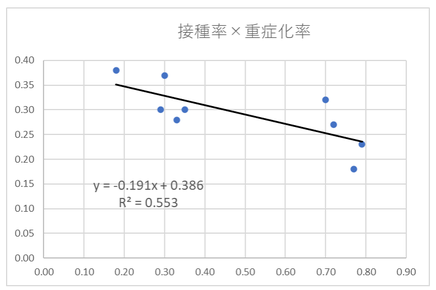

図2．ワクチン接種率と重症化率のグラフと回帰直線

両例とも9のデータの相関係数は -0.743 で，回帰直線は $y = 0.386 - 0.191x$ である．なお，直線の傾き (-0.191) のP値は0.022で統計的に5%有意である．このデータから言えることは，女子学生の比率が大きければ英語試験での不合格率が低いということ，あるいは，ワクチン接種率が大きければ重症化率が低いということである．これを「女子学生の比率を大きくすると英語試験の不合格率が下がる」，「ワクチン接種率を高くすると重症化率が下がる」とか，甚だしくは「女子学生のほうが英語能力が高い」と解釈してはならない．

3．エコロジカルインファレンスの推定結果

3.1. 推定対象

与えられるデータは集計データであるが，我々の本当の興味の対象は個人に関する条件付き確率であろう．具体的には，英語力の例では $p = P$(不合格 | 女子)，$r = P$(不合格 | 男子) であり，ワクチンの例では $p = P$(重症化 | ワクチン接種あり)，$r = P$(重症化 | ワクチン接種なし) である．これらの個人に関するパラメータを集計データから推定することが問題となる．

3.2 Goodman 回帰

集計データから得られた結果を個人の行動に対してむやみに適用してはならないという Robinson (1950) の問題提起の数年後，Leo Goodman は，それを回帰の問題として定式化し，一つの解決法を与えた (Goodman, 1953)．これをエコロジカル回帰あるいは Goodman 回帰 という．推定の詳細は他の文献に譲るとして，結果としては，推定値として $p = 0.195$，$r = 0.386$ が得られる．これより，女子学生のほうが不合格率が低い，ワクチン接種ありのほうが重症化率が低いという結論になる．

3.3. Friedman の近隣モデル

David Freedman は近隣を提唱した (Freedman, 1991)．近隣モデルは，第 i 地域での p_i と r_i は等しく，X_i の線形関数，すなわち $r_i = p_i = a + bX_i$ $(i = 1, \ldots, m)$ としたものである．ここで特に $r_i = p_i$ と置くと $Y_i = r_i$ $(i = 1, \ldots, m)$ となる．近隣モデルは，本来推論の対象で異なった値であるからこそ意味のあるパラメータの p_i と r_i を等しいと置いてしまっていて，例2のワクチンの例では，ワクチン接種者と非接種者で重症化率が同じであるという意味となり，現象の解明につながっているとはとても言えないであろう．しかし，例1の学生の性別と英語力の例ではどうであろうか．この例で，女子学生の比率が高く，英語試験の不合格率も低かったのは英米文学科であって，女子学生の英語力もそれなりに高いが男子学生の英語力も同程度に高いことが容易に想像される．逆に，男女ともに英語試験の成績が悪く不合格率の高かったのは理工学部の某学科であった．すなわちこの例では，近隣モデルの $r_i = p_i$ が現象を的確に表現しているのであり，近隣モデルが現象を表す妥当性を持つ例となっている．

4．おわりに

ここで扱った例では，まったく同じデータでありながら，それが何を意味するものであるかによって適切なモデルが異なり，したがって，推定結果として何を採用し，それをどう解釈すべきかが全く異なるというものである．特にここで扱った集計データでは，背景情報として何を持つのかが決定的に重要であることを示している．統計ソフト（あるいは近年の呼び名ではアプリ）にデータを入れさえすれば結果が得られるというのは全くの幻想である．

エコロジカルインファレンスの研究は，Robinson (1950) の問題提起以来長い歴史を有している（Kousser (2001) を参照）．そして，King (1997) によりその名目は一新した．また，Greenland (2001) あるいは Greenland and Robins (1994) に代表される疫学分野，そしてKing (1997) に代表される政治学分野での応用例が多い．詳細は，King, Rosen and Tanner（2004) および Wakefield (2004) を参照されたい．

この発表は，1冊の書物を必要とする話題を限られたスペースでまとめたものであるが，本来はじっくり考えながら勉強すべき話題である．実際，筆者は単行本にすべく準備中である．

参考文献

岩崎　学 (2019). 事例で学ぶ！あたらしいデータサイエンスの教科書．翔泳社．

岩崎　学 (2022). 今こそ考える「因果」と「相関」．応用統計学会フロンティアセミナー資料（「応用統計学」(2022), **52**（掲載予定））．

Achen, C. H. and Shively, W. P. (1995). *Cross-level Inference*. The Chicago University Press.

Bossarte, R. (2008). *A Contextual Effects Assessment of Ecological Inference*. VDM Verlag Dr. Müller.

Freedman, D. A. (1991). Ecological regression and voting rights. *Evaluation Review*, **15**, 673-711.

Goodman, L. A. (1953). Ecological regression and the behavior of individuals. *American Sociological Review*, **18**, 663-664.

Greenland, S. (2001). Ecologic versus individual-level sources of bias in sociologic estimates of contextual health effects. *International Journal of Epidemiology*, **30**, 1343-1350.

Greenland, S. and Robins, J. (1994). Ecologic studies - biases, misconceptions, and counterexamples (with comment). *American Journal of Epidemiology*, **139**, 747-771.

King, G. (1997). A *Solution to the Ecological Inference Problem*. *Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press.

King, G., Rosen, O. and Tanner, M. A. (2004). *Ecological Inference. New Methodological Strategies*. Cambridge University Press.

Kousser, J. M. (2001). Ecological inference from Goodman to King. *Historical Methods*, **34**, 101-126.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, **15**, 351-357.

Wakefield, J. (2004). Ecological inference for 2 × 2 tables (with discussion). *Journal of the Royal Statistical Society*, *Series A*, **167**, 385-445.