シンプソンのパラドックス

各々の部分データから、これらをまとめた全体へと推論することはよく用いられることである。しかし各部分について成立していることが、合併した全体に当てはまるとは限らない。この現象を指摘した 1951 年の論文、E.H.Simpson;"The interpretation of inerraction in contingency tables", J.Royal Stat Soc, B13:238-241 である。しかし実際はもっと過去に遡って、有名な J.Yule, K.Pearson らも議論しているが、simpson は機知にとんだ意外な解説により、明快に論じたことからその由来があると言われる。たとえば、男性と女性の２つグループで、それぞれに関する新薬の効果を調べる。それぞれのデータの解析結果での結論が、合併したデータのもとでのでは逆の結論となってしまう。こんなことが本当に起こるであろうか。

２×２の分割表 (contingency tables) とは、つぎに与えられるような形:

度数	B_1	B_2	計 (周辺)
A_1	n_{11}	n_{11}	$n_{1.}$
A_2	n_{21}	n_{21}	$n_{2.}$
計 (周辺)	$n_{.1}$	$n_{.2}$	n

という行列形である。たとえば予防注射の効用を調べるとき、属性を A,B の二つでそれぞれがまた 2 つに分ける。A_1: 予防注射を受けた, A_2: 受けない、またその効果について、B_1: 非罹患,B_2: 罹患という形に分類する。4 個の列と行の合計（周辺度数）$n_{1.}$、$n_{2.}$、$n_{.1}$、$n_{.2}$ を与えて、条件：

$$n_{1.} = n_{11} + n_{12}, \quad n_{2.} = n_{21} + n_{22}$$
$$n_{.1} = n_{11} + n_{21}, \quad n_{.2} = n_{21} + n_{22}$$

を満たすような 4 つの数 $(n_{11}, n_{12}, n_{21}, n_{22})$ を求める。ただしセルの総合計を $n = n_{1.} + n_{2.} = n_{.1} + n_{.2}$ とおいた。もしこの属性が独立で、各々の項目に含む確率が属性 A については $p_1, p_2(= 1-p_1)$ として、属性 B では $q_1, q_2(= 1-q_1)$ とし、属性の間には、独立と仮定する。するとこれを満たすような場合の数は、超幾何分布に従う。これからの目的は、周辺度数 $(n_{1.},n_{2.},n_{.1},n_{.2})$ を与えたとき、各セルの数が $(n_{11}, n_{12}, n_{21}, n_{22})$ となる事象の条件付き確率（簡略化して表現して）$p(n_{ij}\,|\,n_{.i}, n_{j.})$, $(i, j = 1, 2)$ を求める。一般に r 行 s 列のデータ $i = 1, 2, \cdots, r, j = 1, 2, \cdots, s$ であっても同様の計算である。条件付き確率に対して $p(n_{ij}\,|\,n_{.i}, n_{j.}) = \dfrac{p(n_{ij})}{p(n_{.i}, n_{j.})}$ を求める。分子には同時確率であるから $p(n_{ij}) = \dfrac{n!}{\Pi_{i,j} n_{ij}!} \left(p_1^{n_{1.}} p_2^{n_{2.}}\right)\left(q_1^{n_{.1}} q_2^{n_{.2}}\right)$ つまり

$$p(n_{11}, n_{12}, n_{21}, n_{22}) = \frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}\; p_1^{n_{1.}} p_2^{n_{2.}} q_1^{n_{.1}} q_2^{n_{.2}}$$

もう一つの確率 $p(n_{i.}, n_{.j})$ は周辺密度として計算する。条件を満たすものにわたって和をとるから、$p(n_{i.}, n_{.j}) = \sum_{*} p(n_{ij})$, ここで和に関する $*$ は条件 $(i, j) : \sum_i n_{ij} = n_{.j}, \sum_j n_{ij} = n_{i.}$ となるすべての n_{ij} についての和を表すとする。実際

$$\begin{aligned} p(n_{i.}, n_{j.}) &= \sum_{*} p(n_{ij}) = \sum_{*} \frac{n!}{\Pi_{i,j} n_{ij}!} \left(p_1^{n_{1.}} p_2^{n_{2.}}\right)\left(q_1^{n_{.1}} q_2^{n_{.2}}\right) \\ &= n! \sum_{*} \frac{1}{\Pi_{i,j} n_{ij}!} \left(p_1^{n_{1.}} p_2^{n_{2.}}\right)\left(q_1^{n_{.1}} q_2^{n_{.2}}\right) = \frac{(n!)^2}{n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!} \left(p_1^{n_{1.}} p_2^{n_{2.}}\right)\left(q_1^{n_{.1}} q_2^{n_{.2}}\right) \end{aligned}$$

となる。なぜならば、多項式の展開 $(x_1+x_2)^{n_{.1}}(x_1+x_2)^{n_{.2}} = (x_1+x_2)^{n_{.1}+n_{.2}} = (x_1+x_2)^n$ において係数 $x_1^{n_{1.}} x_2^{n_{2.}}$ を比較すると $\sum_{*} \dfrac{n_{.1}!}{n_{11}!\, n_{21}!} \dfrac{n_{.2}!}{n_{12}!\, n_{22}!} = \dfrac{n!}{n_{1.}!\; n_{2.}!}$ であるから, $\sum_{*} \dfrac{1}{n_{11}!\, n_{21}!\, n_{12}!\, n_{22}!} = \dfrac{n!}{n_{1.}!\, n_{2.}!\, n_{.1}!\, n_{.2}!}$ を得

る。したがって

$$p(n_{ij}\,|\,n_{i.},n_{.j}) = \frac{(\Pi_i n_{i.}!)(\Pi_j n_{.j}!)}{n!\,\Pi_{i,j} n_{ij}!}, \qquad \forall i,j$$

$$p(n_{11},n_{12},n_{21},n_{22}\,|\,n_{1.},n_{2.},n_{.1},n_{.2}) = \frac{1}{n_{11}!\,n_{12}!\,n_{21}!\,n_{22}!}\left(\frac{n_{1.}!\,n_{2.}!\,n_{.1}!\,n_{.2}!}{n!}\right)$$
$$= (n_{ij}\text{によらない定数})\frac{1}{n_{11}!\,n_{12}!\,n_{21}!\,n_{22}!}$$

上で述べた条件を満たすすべての $(n_{11},n_{12},n_{21},n_{22})$ を加え合わせれば、1 となるものが（n_{ij} によらない定数) である。

つぎのデータは仮想的なものであるが、2 つの集団での効果（比率）は、合併されたものとは反対の効果（比率）となっている。

男性	効果あり	なし	計
属性 A	$a=3$	4	$b=7$
属性 $\overline{A}$	$c=1$	1	$d=2$

女性	効果あり	なし	計
属性 B	$p=1$	5	$q=6$
属性 $\overline{B}$	$r=1$	4	$s=5$

$$\frac{3}{7}=\frac{a}{b}<\frac{c}{d}=\frac{1}{2}$$

$$\frac{1}{6}=\frac{p}{q}<\frac{r}{s}=\frac{1}{5}$$

$$\Rightarrow$$

合併集団	あり	なし	計
属性 $A\cup B$	4	9	13
属性 $\overline{A\cup B}$	2	5	7

$$\frac{a+p}{b+q}=\frac{4}{13}>\frac{2}{7}=\frac{c+r}{d+s}$$

この分割表では、男性と女性と部分データの結論：効果（比率）はそれぞれのデータとも、「効果あり」は認められないが、合併した集団データに対しての効果（比率）では、逆転が起こっている。

この現象は行列式の非加法性に対応することがわかる。なぜなら、行列 $X=\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ とおくと、条件 $\frac{a}{b}<\frac{c}{d}$ は、行列式 $\det|X| = ad-bc<0$ と同じ、ただし成分は 0 ではないとする。したがって、同様に行列 $Y=\begin{pmatrix} p & q \\ r & s \end{pmatrix}$ とおき、合併したデータは、行列 $X+Y=\begin{pmatrix} a+p & b+q \\ c+r & d+s \end{pmatrix}$ である。しかし 2 つの条件：$\det|X|<0, \det|Y|<0$ からは、行列式の加法性が成り立たないから、$\det|X+Y|<0$ とはならない。この場合が simpson のパラドックスに対応している。