レコードリンケージにおける統計モデルによる個人同定処理の自動化について

広島大学原爆放射線医科学研究所計量生物分野 佐 藤 健 一

広島大学原爆放射線医科学研究所放射線分子疫学分野 早 川 式 彦

広島大学原爆放射線医科学研究所国際放射線情報センター 隅 田 治 行

広島大学原爆放射線医科学研究所計量生物分野 大 瀧 慈

要 旨 レコードリンケージにおける個人同定処理は、計算機が未発達であった時期から必要とされていたこともあり、国内では経験的な手法についての議論が先行している。本稿では、レコード値の一致型ペア数に対して、2項分布の和を想定し、ポアソン近似することで統計モデルに基づいた手法を議論する。また、経験的な知識の統計モデルへの利用方法についても検討し、計算機による同定処理の自動化を行う際に有用なレコード値の頻度を考慮した判定基準を提案した。解析例によって、統計モデルの適合度を検証することができ、新規でレコードリンケージを行う場合のみならず、既存の照合結果の検証にも有効であることが示唆された。

1. はじめに

同一人の情報を含む異なる2つのデータファイルがあるときに、同一人に関するレコードの結合を考える。日本ではアメリカの Social Security Number のような共通の個人識別番号がないため、多くの場合、両方のデータにある共通項目、例えば、氏名、住所、生年月日など、を手掛かりに個人を同定する必要がある。このような問題は Dunn(1946)によってレコードリンケージとして提起された。その後、Newcombe, et al.(1959), Newcombe and Kennedy(1962)は、2つのファイルから取り出したレコードペアの各フィールド値の一致型に着目し、一致型別のレコードペア数を同一人に関するレコードペア数と別人に関するレコードペア数の和として統計モデルを構築した。このモデルは、現在にいたるまで標準的に利用されている。Fellegi and Sunter(1969)および Tepping(1968)は、このモデルのもとで決定論の枠組みを構築し、与えられたレコードのペアを、同一人に関するレコードとすべきか、別人に関するレコードとすべきかを判断する検定方式について議論した。D'Andrea and Bois(1969)は、レコード値の情報に欠損値がある場合も論じている。一方、日本においては、計算機を利用した実地的、経験的な側面から、柳川ら(1971)、小野ら(1976)、Ohshima, et al.(1979)、近藤ら(1979)、によって、疫学調査における照合作業についての議論がなされており、大瀧ら(1982, 1985)によって、統計モデルを利用した個人同定処

理自動化の試みが行われた。近年では、レコードリンケージのためのソフトウェア、Automatch (Matchware Technologies, Inc., Silver Spring, MD)が開発されており、そのアルゴリズムの解説はJaro(1995)に、使用報告として、例えば、Newman and Brown(1997)がある。

本稿では、Newcombe, et al.(1959)およびNewcombe and Kennedy(1962)が提案した、統計モデルに基づくレコードリンケージを第2節に説明し、第3節では経験的な情報をモデルに取り入れるための確率構造について触れる。第4節において、同一人に関するレコードペアであることを判定するための基準として、Fellegi and Sunter(1969)およびTepping(1968)が議論した基本統計量を紹介する。さらに、この統計量を、レコード値の頻度を反映するように発展させ、新たに、固定されたレコードごとの基準を提案する。第5節においてはモデルの推定について、経験的な観点から素朴な推定量を含め議論する。第6節では実際のデータを通して、レコードリンケージの手順を紹介し、第7節において、実用上の問題点を議論しながら統計モデルに基づくレコードリンケージの有用性を検証する。

2. 数理モデル

まず、Newcombe, et al.(1959), Newcombe and Kennedy(1962)が提案したレコードリンケージのための統計モデルを紹介する。既存のデータファイルをA, 照合したい新しいデータファイルをBとし、レコード値はK個の共通な項目から成るとする。このとき、データファイルAおよびBはレコードの集合として、それぞれ、 $\mathcal{A} = \{a_i | i = 1, \dots, N_A\}$, および $\mathcal{B} = \{b_i | i = 1, \dots, N_B\}$, と表現できる。ただし、 $a_i = (a_{i1}, \dots, a_{iK})$, $b_i = (b_{i1}, \dots, b_{iK})$. ここで、 $\mathcal{A}$ および $\mathcal{B}$ は所与の有限母集団とする。両方のファイルからレコードのペア (a, b) を取ってくると、表1に例を挙げる各項目値の一致状況を示す一致型、

$$t(a, b) = (t_1, \dots, t_K), \quad (1)$$

を得る。ただし、 $k = 1, \dots, K$ に対して、

$$t_k = \begin{cases} 1, & a_k = b_k \text{ の場合,} \\ 0, & a_k \neq b_k \text{ の場合,} \end{cases}$$

とする。

表1. レコードペアと一致型の例.

レコード値と一致型	氏	名	元号	年	月	日
a	佐藤	健一	明治	4	4	9
b	佐藤	春夫	明治	4	4	9
$t(a, b)$	1	0	1	1	1	1

レコードペアを1つ取ってくれば、それは同一人に関するレコードのペア(以下、同一人ペアと呼ぶ)か、別人に関するレコードのペア(別人ペア)のどちらかである。そして、一般的には、一致型の成分に0が多ければ別人ペアである可能性が高く、逆に1が多ければ同一人ペアである可能性が高いと予想される。ここでは、より定量的な評価が行えるように統計モデルによる記述を議論する。

すべてのペア数における同一人ペア数の相対頻度 $\xi \in [0, 1]$ を用いて、同一人ペア数および別人ペア数は、それぞれ、 $N^{(S)} = \min(N_A, N_B)\xi$ 、および $N^{(D)} = N_A N_B - N^{(S)}$ 、同一人ペアおよび別人ペアにおいて一致型 t が観測される確率を、それぞれ $p^{(S)}(t)$ および $p^{(D)}(t)$ とすると、一致型 t を持つペア数 $y(t)$ に対して、以下の独立な 2 項分布の和が想定できる。

$$y(t) = y^{(S)}(t) + y^{(D)}(t), \quad y^{(j)}(t) \sim B(N^{(j)}, p^{(j)}(t)), \quad j \in \{S, D\}, \quad (2)$$

ただし、 $B(N, p)$ は試行回数 N 、成功確率 p を持つ 2 項分布を表し、 $y^{(S)}(t)$ および $y^{(D)}(t)$ は、それぞれ一致型 t を持つ同一人ペア数および別人ペア数である。したがって、(2) 式の期待値は、

$$\lambda(t) = \lambda^{(S)}(t) + \lambda^{(D)}(t), \quad \lambda^{(j)}(t) = N^{(j)} p^{(j)}(t), \quad j \in \{S, D\}, \quad (3)$$

とかける。レコードリンケージの手法が必要とされる状況では、 $N^{(j)}$, $j \in \{S, D\}$ が十分大きく、 $p^{(j)}$, $j \in \{S, D\}$ は十分小さい、ことが想定できるので、2 項分布を、それぞれポアソン分布で近似する、すなわち、 $y^{(j)}(t) \sim \text{Poisson}(\lambda^{(j)}(t))$, $j \in \{S, D\}$ 、および、 $y(t) \sim \text{Poisson}(\lambda(t))$ 、とする。ただし、 $\text{Poisson}(\lambda)$ は平均 λ を持つポアソン分布を表す。したがって、 $y(t)$ が与えられた下での同一人ペア数 $y^{(S)}(t)$ の分布は、

$$y^{(S)}(t) \sim B(y(t), r^{(S)}(t)), \quad r^{(S)}(t) = \frac{\lambda^{(S)}(t)}{\lambda(t)}, \quad (4)$$

とかける。

共通項目数が K 個ある場合、とり得る一致型の場合の数は、すべての項目が不一致の場合からすべての項目が一致まで、 2^K 通りあり、 $\{t_m | m = 1, \dots, 2^K\}$ と記す。一方、全ペア数は $N_A N_B$ と固定されているため、次の (5) 式が成り立つ。

$$\sum_{m=1}^{2^K} y(t_m) = N_A N_B. \quad (5)$$

それゆえ、すべての一致型別ペア数に独立なポアソン分布を仮定することは適切ではないので、同一人ペアが多く含まれると思われる、ある程度の高い一致状況をもつ一致型を $M (\ll 2^K)$ 個選び、添え字を書き直すことによって、その尤度関数は近似的に、

$$L(y(t_1), \dots, y(t_M) | \lambda(t_1), \dots, \lambda(t_M)) = \prod_{m=1}^M \frac{\lambda(t_m)^{y(t_m)} e^{-\lambda(t_m)}}{y(t_m)!}, \quad (6)$$

とかける。

3. 項目間の確率構造の独立性

同一人ペアおよび別人ペアにおいて一致型 t が観測される確率について考える。同一人ペアにおいて、対応する項目が一致しない原因としては、同一人に関するレコードが異なる時期に、 $\mathcal{A}$ および $\mathcal{B}$ に登録された場合などがある。例えば、既婚なら 1、未婚なら 0 をとる、「既婚」という項目があるとする。特に、日本においては結婚後に、女性が苗字を変えることが多く、同一人ペアであっても、「既婚」値が不一致なら「苗字」値が不一致になる確率は高い。したがって、「既婚」

と「苗字」の項目が一致する確率は独立でなく、一般的には、項目間の独立性を仮定できない。また、容易に分かるように、別人ペアにおいては、「既婚」値と「苗字」値が一致する確率は独立として扱えるため、同一人ペアと別人ペアでは、項目間における独立性に異なる構造が必要となる。

そこで、同一人ペア内および別人ペア内でそれぞれ K 個の項目を重複なく、いくつかの項目グループに分け、グループ間の独立性を次のように仮定する。

$$p^{(j)}(\mathbf{t}) = \prod_{l=1}^{L^{(j)}} p_l^{(j)}(\mathbf{t}_l^{(j)}), \quad \mathbf{t} = (\mathbf{t}_1^{(j)}, \dots, \mathbf{t}_{L^{(j)}}^{(j)}), \quad j \in \{S, D\}, \quad (7)$$

ただし、 $\mathbf{t}_l^{(j)}$ は長さ $k_l^{(j)}$ のベクトルで、 $\sum_{l=1}^{L^{(j)}} k_l^{(j)} = K$ を満たす。ここで、第 l グループの項目数を k_l 、その一致型を $\mathbf{t}_l$ とすると、その確率は $2^{k_l} - 1$ 個のパラメータを持つ多項分布で記述できる。

さらに、項目グループの最適化、使用する項目の選択などを議論することもできるが、これらは、対象となるデータベースの担当者によって経験的な側面から議論されるべき問題かも知れない。

例えば、地方の住民が登録されるデータベースであれば、「住所」と「苗字」に関連性がある場合もあり、登録者についての付加的な情報が有用となる。

4. 同一人ペア確率

我々の目的は、与えられたレコードのペアが「同一人に関するレコードである」ことの確からしさを定量的に評価し、これに基づいてレコードリンケージを行うことにある。一致型 $\mathbf{t}$ を持つペア集合から無作為に 1 つ取り出したレコードのペアが同一人ペアである確率(以後、同一人ペア確率)は、(4) 式で与えられ、次式のように表すことができる。この確率が高い一致型を持つレコードペアが同一人ペアの候補となる。

$$r^{(S)}(\mathbf{t}) = \frac{N^{(S)} p^{(S)}(\mathbf{t})}{N^{(S)} p^{(S)}(\mathbf{t}) + (N_A N_B - N^{(S)}) p^{(D)}(\mathbf{t})}. \quad (8)$$

Fellegi and Sunter (1969) および Tepping (1968) では、オッズ比型、 $r^{(S)}(\mathbf{t}) / \{1 - r^{(S)}(\mathbf{t})\}$ も議論されている。

これまでの議論は、すべてのレコードペアを考えたものであったが、新しく得られたファイルの中の固定されたレコードと同一人ペアになるものを既存のファイルの中から探すという状況がより現実的である。そこで、ファイル B のレコードを一つ固定し、ファイル A の各レコードとの同一人ペア確率を考える。簡単のため、1) データファイル A に同一人に関する重複レコードはなく、2) $N_A \geq N_B$ とすれば、レコード $\mathbf{b}$ と同一人に関するレコードペアは、ファイル A の中に、 $N^{(S)}/N_B$ ペア存在することが期待される。

同一人ペアに関する確率は、記入ミス、転記ミス、経年変化などレコード値そのものにはほとんど依らない構造を持つことが経験的に知られているが、別人ペアに関する確率は、そのレコード値の頻度に大きく依存する。そこで、レコード $\mathbf{b}$ がデータファイル A の別人に関するレコードと一致型 $\mathbf{t}$ を持つ確率を $p^{(D)}(\mathbf{t}|\mathbf{b})$ とすれば、ファイル B のレコード $\mathbf{b}$ を一つ固定したときの、同一人ペア確率は、 $\xi = N^{(S)}/N_B$ を用いて、

$$r^{(S)}(\mathbf{t}|\mathbf{b}) = \frac{\xi p^{(S)}(\mathbf{t})}{\xi p^{(S)}(\mathbf{t}) + (N_A - \xi) p^{(D)}(\mathbf{t}|\mathbf{b})}, \quad (9)$$

とかける．一見すると，別人ペアである確率のみがレコード値に依存するように見えるが，例えば，珍しい名前で一致している場合，別人ペアで一致する確率は小さくなり，相対的に，同一人ペア確率は高くなる．

実際の照合処理においては，同一人ペア確率を昇順もしくは降順に並び替えることにより，自動判定処理，もしくは，人手処理の対象とすべき複数のレコードペアを選出するのが一般的である．今，すべてのレコードペアに対する (8) 式もしくは (9) 式による同一人ペア確率を $r_1^{(S)}, \dots, r_{N_A N_B}^{(S)}$ と記し，同一人ペア確率がある区間 $I \subseteq [0, 1]$ に含まれる場合の添え字集合， $J(I) = \{i | r_i^{(S)} \in I, i = 1, \dots, N_A N_B\}$ を考える．このとき，区間 I に含まれる同一人ペア数および別人ペア数の期待値は，それぞれ，

$$q^{(S)}(I) = \sum_{i \in J(I)} r_i^{(S)}, \quad q^{(D)}(I) = \sum_{i \in J(I)} (1 - r_i^{(S)}), \quad (10)$$

と記述できる．区間 I に含まれるレコードペア集合を同一人ペアとして判定する場合には，その中に含まれる別人ペア数 $q^{(D)}(I)$ が，逆に，別人ペアとして判定する場合には同一人ペア数 $q^{(S)}(I)$ が，それぞれの誤判別の指標となる．照合処理の担当者は，これらの統計量を利用することで照合処理における精度を逐次把握することができ，処理精度の制御が可能となる．

同一人ペア確率を (8) 式および (9) 式で与えたが，すべてのレコードペアの集合を考えた場合，どちらの基準の総和も同一人ペア数 $N^{(S)}$ を表している．したがって，以下の等式が近似的に成り立つ．

$$\frac{\sum_{(a,b) \in \mathcal{A} \times \mathcal{B}} r^{(S)}(t(a,b))}{\sum_{b \in \mathcal{B}} \sum_{a \in \mathcal{A}} r^{(S)}(t(a,b)|b)} = 1. \quad (11)$$

さらに，任意の一致型 t_0 を持つレコードペアの集合に対して，

$$\frac{r^{(S)}(t_0)}{\sum_{(a,b) \in C(t_0)} r^{(S)}(t(a,b)|b)} = 1,$$

を考えることもできる．ただし， $C(t_0) = \{(a,b) \in \mathcal{A} \times \mathcal{B} | t(a,b) = t_0\}$ である．

5. 未知パラメータの推定

簡単のため，同一人ペアおよび別人ペアそれぞれについて一致型の各成分が独立な場合を考える．すなわち，

$$p^{(j)}(t) = \prod_{k=1}^K p_k^{(j)}(t_k), \quad p_k^{(j)}(t_k) = (\mu_k^{(j)})^{t_k} (1 - \mu_k^{(j)})^{1-t_k}, \quad j \in \{S, D\}, \quad (12)$$

ただし， $\mu_k^{(S)}$ および $\mu_k^{(D)}$ は，それぞれ，同一人ペアおよび別人ペアにおいて，第 k 項目が一致する確率を表す．独立でない場合の議論は，(12) 式を多項分布を用いて表記すればよい．推定すべきパラメータの数が多いため，安定した推定値を得るために，ここでは 2 段階の推定方法を議論する．場合によっては，同時に推定することも可能である．まず，別人ペアにおいて一致型 t が観察される確率を考える． $\mathcal{A} \times \mathcal{B}$ に属するレコードペアにおいて，第 k 項目値が一致する頻度を $f_k^{(D)}$ とすれば，別人ペア数に比べて同一人ペア数が少ないことから，(12) 式の $\mu_k^{(D)}$ の素朴な推定値は $f_k^{(D)} / (N_A N_B)$ で与えられる．次に， $\mu_k^{(D)}$, $k = 1, \dots, K$ を既知パラメータとし， M 個のペア数データ $\{y(t_m) | m = 1, \dots, M\}$ から，未知パラメータ ξ および $\mu_k^{(S)}$, $k = 1, \dots, K$ を (6) 式の尤度を

用いて最尤推定する．このとき，過去の照合による各パラメータの事前情報があれば，最適化の初期値として利用できる．未知パラメータが推定されれば， $\hat{\xi}$ および $\hat{p}^{(j)}(t)$, $j \in \{S, D\}$ を用いて， $\hat{\lambda}^{(j)}(t) = \hat{N}^{(j)} \hat{p}^{(j)}(t)$, $j \in \{S, D\}$, $\hat{N}^{(S)} = \min(N_A, N_B) \hat{\xi}$, $\hat{N}^{(D)} = N_A N_B - \hat{N}^{(S)}$, のように推定でき，(8) の同一人ペア確率が推定できる．

同様に，レコード $\mathbf{b}$ を固定したときの同一人ペア確率の推定については， $\mathcal{A} \times \{\mathbf{b}\}$ に属するレコードペアにおいて，第 k 項目値が一致する頻度を $f_k^{(D)}(\mathbf{b})$ とすると，同一人ペアの存在を無視することで $\mu_k^{(D)}(\mathbf{b})$ は $f_k^{(D)}(\mathbf{b})/N_A$ によって推定できる．このとき，定義から，

$$\frac{f_k^{(D)}}{N_A N_B} = \frac{1}{N_B} \sum_{\mathbf{b} \in \mathcal{B}} \frac{f_k^{(D)}(\mathbf{b})}{N_A}, \quad (13)$$

が成り立つことに注意する．なお， $p^{(D)}(t|\mathbf{b})$ は，その定義から，項目間の独立性を仮定しない場合でも，ファイル A における一致型 t の割合として推定可能であるが，観測頻度が少ない場合には推定精度が悪くなるため，実用的ではない．また，第 6 節において後述するが，実際には (11) 式を収束判定基準として，上記の手続きを繰り返すことで，推定値を改良する．

6. 実データを通した照合処理手順の紹介

当研究所には，現在 288,266 人の被爆者データベース(ファイル A)があり，被爆者の死因と被爆線量などとの関連性について研究がなされている．しかしながら，死亡情報(ファイル B)は，他の機関が保有しており，毎年，ファイル A および B において，同一人に関するレコードがあるかどうか経験的な手法をもとに調べている．

ここでは，統計モデルによるレコードリンケージの解析例として，1997 年の照合済みのファイルを利用し，その有効性を検証する．同一人のペアにおける性別の不一致については，既に人手によって修正されており，また，男女に共通のパラメータを仮定することは現実的ではないため，ここでは男女別々に解析を行うこととした．男性については， $N_A = 138,594$ および $N_B = 12,454$ ，また，経験的な方法による確認済みの同一人ペア数は $N^{(S)} = 1,797$ ，女性については， $N_A = 149,677$ ， $N_B = 10,542$ および $N^{(S)} = 1,889$ であった．

なお，確認済みのペア数は照合費用の都合により真の同一人ペア数よりも低く同定されていることが認識されている．共通項目としては，漢字表記による氏名(氏，名)，元号を含む生年月日(元号，年，月，日)を用いた．住所は重要な項目であるが，区政の前後で表記が異なるものが多いであったため解析例には用いなかった．

独立モデル (12) を仮定した場合，レコードリンケージの手順は以下のようになる．

- 1) すべてのレコードペアについて一致型を調べ(表 1 参照)，一致型別ペア数表を作成する(参照：表 2.1 および 2.2)．このとき，第 k 項目が一致する頻度 $f_k^{(D)}$, $k = 1, \dots, K$ を数える(表 3 参照)．
- 2) 一致型別ペア数表から，一致型の成分に 1 が少ないもの，もしくはペア数が多いものを取り除く．
- 3) 抜粋した一致型別ペア数表に，モデル (12) を独立に仮定し，(6) 式の尤度を用いて推定値 $\hat{\xi}$ および $\hat{\mu}_k^{(j)}$, $k = 1, \dots, K$, $j \in \{S, D\}$ を求める．このとき， $\mu_k^{(D)}$ を既知とし， $f_k^{(D)}/N_A N_B$ を与

- えることで、未知パラメータを減らすことも可能である(表 3 参照).
- 4) 一致型別に, $\hat{\lambda}^{(j)}, j \in \{S, D\}$ の推定値を求め, ペア数の予測値 $\hat{y} = \hat{\lambda}$, 同一人ペア確率 $\hat{\rho}^{(S)}$, 同一人ペア数の予測値 $\hat{y}^{(S)}$ を計算し, 観測数 y , 経験的に確認されていれば $y^{(S)}$ との適合を見る(参照: 表 2.1 および 2.2).
 - 5) 次に, 固定されたレコード値 $\mathbf{b}$ についてレコードリンクージュを行うために, ファイル A におけるレコード値の頻度 $f_k^{(D)}(\mathbf{b}), k = 1, \dots, K$ を調べ(表 4 参照), $\hat{\rho}_k^{(D)}(\mathbf{b}) = f_k^{(D)}(\mathbf{b})/N_A, k = 1, \dots, K$ とし, 3) で推定された $\hat{\rho}_k^{(S)}, k = 1, \dots, K$ および $\hat{\xi}$ を用いて, (9) 式の同一人ペア確率を算出する.
 - 6) すべてのレコードペアに含まれる同一人ペア数 $N^{(S)}$ に関する等式 (11) が満たされていれば終了し, 満たされていない場合は, 5) で算出された同一人ペア確率の総和を改めて $\hat{N}^{(S)}$ とおき, 3) の同一人ペア数の相対頻度を $\hat{\xi} = \hat{N}^{(S)} / \min(N_A, N_B)$ と与えた後に, 手順 3)–5) を再度行う.
 - 7) 誤判別に関する指標 (10) 式をもとに, 自動判定および人手判定に該当するレコードペアを選出する(表 5 参照).

表 2.1. 男性における一致型別ペア数データと統計モデルによる予測値. すべての一致型 $2^K, K = 6$ から, 一致型の成分に 0 が高々 2 つのものを抜粋した. 統計モデルの推定には $m = 1, \dots, 7$ までのデータだけを使用した.

m	一致型						ペア数		同一人ペア確率		同一人ペア数	
	氏	名	元	年	月	日	実測値 ¹⁾	予測値 ²⁾	実測値 ³⁾	予測値 ⁴⁾	実測値 ⁵⁾	予測値 ⁶⁾
1	1	1	1	1	1	1	1,418	1452.29	99.859	99.996	1,416	1417.94
2	1	1	1	1	1	0	46	47.55	93.478	96.724	43	44.49
3	1	1	1	1	0	1	2	2.07	100.000	69.533	2	1.39
4	1	1	1	0	1	1	10	10.35	90.000	81.460	9	8.15
5	1	1	0	1	1	1	4	4.15	100.000	96.938	4	3.88
6	1	0	1	1	1	1	304	313.10	48.684	73.288	148	222.80
7	0	1	1	1	1	1	237	244.21	48.101	82.467	114	195.45
8	1	1	1	1	0	0	43	16.06	13.953	0.284	6	0.12
9	1	1	1	0	1	0	63	48.99	1.587	0.545	1	0.34
10	1	1	1	0	0	1	26	19.73	0.000	0.042	0	0.01
11	1	1	0	1	1	0	3	3.35	0.000	3.797	0	0.11
12	1	1	0	1	0	1	1	1.31	0.000	0.304	0	0.00
13	1	1	0	0	1	1	5	3.99	0.000	0.584	0	0.03
14	1	0	1	1	1	0	3,553	2131.63	0.113	0.341	4	12.11
15	1	0	1	1	0	1	1,417	860.01	0.000	0.026	0	0.37
16	1	0	1	0	1	1	2,774	2616.64	0.072	0.051	2	1.41
17	1	0	0	1	1	1	113	173.76	0.885	0.365	1	0.41
18	0	1	1	1	1	0	2,854	1093.93	0.210	0.583	6	16.64
19	0	1	1	1	0	1	1,035	440.36	0.097	0.045	1	0.47
20	0	1	1	0	1	1	1,671	1340.06	0.180	0.087	3	1.46
21	0	1	0	1	1	1	64	89.19	0.000	0.625	0	0.40
22	0	0	1	1	1	1	106,227	58411.58	0.028	0.054	30	57.87

ただし, 実測値および予測値の式は, それぞれ, ペア数については 1) y および 2) $\hat{y}$, 同一人ペア確率については 3) $10^2 y^{(S)} / y$ および 4) $10^2 \hat{\rho}^{(S)}$, 同一人ペア数については 5) $y^{(S)}$ および 6) $\hat{y}^{(S)}$ で与えられる.

表 2.2. 女性における一致型別ペア数データと統計モデルによる予測値。すべての一致型 2^K , $K = 6$ から、一致型の成分に 0 が高々 2 つのものを抜粋した。統計モデルの推定には $m = 1, \dots, 4, 6, 7$ および $m = 11$ のデータだけを使用した。

m	氏 名 元 年 月 日	ペア数		同一人ペア確率		同一人ペア数	
		実測値 ¹⁾	予測値 ²⁾	実測値 ³⁾	予測値 ⁴⁾	実測値 ⁵⁾	予測値 ⁶⁾
1	1 1 1 1 1 1	1,163	1181.07	99.828	99.992	1,161	1162.91
2	1 1 1 1 1 0	67	65.66	82.090	96.465	55	64.63
3	1 1 1 1 0 1	9	9.33	88.889	89.729	8	8.08
4	1 1 1 0 1 1	19	19.70	68.421	83.950	13	15.95
5	1 1 0 1 1 1	0	104.52	0.000	99.807	0	0.00
6	1 0 1 1 1 1	531	545.02	68.173	85.614	362	454.61
7	0 1 1 1 1 1	391	402.23	36.829	83.692	144	327.24
8	1 1 1 1 0 0	104	23.66	6.731	1.898	7	1.97
9	1 1 1 0 1 0	181	77.46	4.420	1.145	8	2.07
10	1 1 1 0 0 1	75	31.73	0.000	0.369	0	0.28
11	1 1 0 1 1 0	4	10.49	0.000	53.320	0	2.13
12	1 1 0 1 0 1	1	2.76	0.000	26.779	0	0.27
13	1 1 0 0 1 1	9	8.13	11.111	17.963	1	1.62
14	1 0 1 1 1 0	3,601	1923.98	0.528	1.301	19	46.84
15	1 0 1 1 0 1	1,504	787.35	0.199	0.420	3	6.32
16	1 0 1 0 1 1	2,764	2593.10	0.543	0.252	15	6.96
17	1 0 0 1 1 1	103	206.65	0.000	19.945	0	20.54
18	0 1 1 1 1 0	5,260	1606.73	0.095	1.124	5	59.10
19	0 1 1 1 0 1	2,054	658.32	0.097	0.363	2	7.45
20	0 1 1 0 1 1	3,541	2168.65	0.141	0.217	5	7.70
21	0 1 0 1 1 1	94	168.14	0.000	17.685	0	16.62
22	0 0 1 1 1 1	111,472	53794.59	0.056	0.247	62	275.62

ただし、実測値および予測値の式は、それぞれ、ペア数については 1) y および 2) $\hat{y}$, 同一人ペア確率については 3) $10^2 y^{(S)} / y$ および 4) $10^2 \hat{y}^{(S)}$, 同一人ペア数については 5) $y^{(S)}$ および 6) $\hat{y}^{(S)}$ で与えられる。

表 3. 未知パラメータの推定値。表 2.1 および 2.2 で与える一致型別ペア数データに対して、独立モデル (12) を仮定した。ただし、男性については、 $\hat{\xi} = 0.16015$, $\hat{N}^{(S)} = 1,994.56$, $N^{(S)} = 1,797$, $N_A = 138,594$, $N_B = 12,454$, 女性については、 $\hat{\xi} = 0.23518$, $\hat{N}^{(S)} = 2,479.26$, $N^{(S)} = 1,889$, $N_A = 149,677$, $N_B = 10,542$.

性別	場合の数と推定値	氏 名 元号 年 月 日
男性	$(N_A N_B) / f_k^{(D)}$	699.02 1364.47 3.07 32.27 11.28 26.40
	$\hat{p}_k^{(S)}$	0.87824 0.86360 0.99725 0.99423 0.99901 0.96931
女性	$(N_A N_B) / f_k^{(D)}$	685.42 819.09 3.11 33.99 11.00 25.22
	$\hat{p}_k^{(S)}$	0.77818 0.71679 0.91884 0.98619 0.99296 0.94910

表 4. 固定されたレコード値とファイル A における頻度の例。ファイル A における頻度は男性のみを対象とし、 $N_A = 138,594$.

レコード値と頻度	氏 名 元号 年 月 日
b	佐藤 春夫 明治 4 4 9
$N_A / f_k^{(D)}(b)$	307.99 969.19 2.84 18.77 13.21 42.9
b	谷崎 潤一郎 明治 19 7 24
$N_A / f_k^{(D)}(b)$	2887.38 138594.00 2.84 40.58 13.91 41.43

表 5. 誤判別の設定と人手判定の対象となるレコードペア数. ファイル A およびファイル B は、ともに男性のみを対象とし、ファイル B のすべてのレコードに対して、ファイル A のすべてのレコードとの (9) 式の同一人ペア確率を算出した. 同一ペア確率に関する区間 $I^\alpha = [\alpha, 1]$ および $I_\beta = [0, \beta]$ を考え、(10) 式で与えられる区間内の別人ペア数 $q^{(D)}(I^\alpha)$ および同一人ペア数 $q^{(S)}(I_\beta)$ を与えた場合の区間 $I_\beta^\alpha = [\beta, \alpha]$ に含まれるレコードペア数を算出した.

同一人ペア候補に含まれる 上側累積別人ペア数	別人ペア候補に含まれる 下側累積同一人ペア数			
	1	10	50	100
1	1,191,271	120,155	16,916	5,687
100	1,190,885	81,723	16,530	121

7. 実データを通した照合処理手順に関する議論

ここでは、6 章で紹介した照合処理をより詳しく議論する. 独立モデル (12) を推定するには、表 2.1 および 2.2 に示すような一部の一致型別ペア数データで十分である. それぞれの一致型に対して (8) 式で与えられる同一人ペア確率(%)を $100 \times \hat{p}^{(S)}$ として与えた. 例えば、表 1 のレコードがともに男性に属する場合、表 2.1 の $m = 6$ の一致型に該当し、このときの同一人ペア確率は 73.288% であり、同じ一致型を持つ 304 のレコードペアの中に、同一人ペアは 222.80 ペアあることが予測され、実際に人手判定によって同定済みのペア数は 148 ペアとなる.

一致型ペア数表から推定したいのは、同一人ペアに関するパラメータであるため、別人ペアに関するパラメータについてはデータベースから算出可能な値、 $\hat{\mu}_k^{(D)} = f_k^{(D)} / (N_A N_B)$ を既知として用いた. 表 3 にその既知パラメータおよび未知パラメータの推定結果を示す. 表 3 において、 $f_k^{(D)} / (N_A N_B)$ の逆数を示しているが、この値は、その項目値の場合の数に相当することに注意する. 統計モデルの推定には表 2.1 および 2.2 の一部のみを使用し、使用しなかった一致型については、予測精度の検討に利用した. 男性ではモデルによる同一人ペア数の予測値 $\hat{y}^{(S)}$ は、確認済みの同一人ペア数 $y^{(S)}$ に対して概ね適合していたが、女性については、観測ペア数と実測値の差異が男性に比べて大きく、同一人ペア数についても適合度が低い結果となった. なお、性別に分けずに一致型レコードペア数を調べると、性が不一致、かつ、他の項目がすべて一致、となるレコードペア数は 2 件のみだった. これは、性別の表 2.1 および表 2.2 の $m = 1$ のペア数に比べてとても少ないため、例に挙げたデータベースにおいては、性別に解析を行っても同一人ペアの取りこぼしは少ないと思われる. ただし、一般的には同一人ペアの中には性別が異なるペアも多く存在する.

固定されたレコードに対する同一人ペア確率は、表 4 に示すレコード値の頻度が基礎となる. 頻度については (13) 式が成り立つので、平均的には、表 3 で与える場合の数に一致する. 例えば、名の項目値に注目すれば、別人ペアにおいて名が偶然一致する確率は平均的には約 $(1364.47)^{-1}$ であるが、「春夫」という名で一致する確率は $(969.19)^{-1}$ となり、高頻度であることがうかがえる. したがって、別人ペアにおいても名が「春夫」で一致する確率は高いと言える. 逆に、別人ペアにおいて名が「潤一郎」で一致する確率は低く、もし一致していれば、同一人ペアである確率が高くなる場合がある.

人手によって照合を行う際には、過去の資料に遡って他の項目、例えば住所の履歴なども使用

しているため、モデルによる予測値がかなり低い場合でも同定が行われているケースがあった。一方で、この解析例においても、統計モデルに基づく同定によって少なくとも2例の同一人ペアが新たに確定した。いずれの場合も、(9)式の同一人ペア確率で発見されており、これらは $p^{(S)}$ はさほど高くないが、別人ペアにおける確率 $p^{(D)}$ が低いために同一人ペア確率が高くなっていた。以上のことから、提案手法である固定レコードに基づくレコードリンケージは新規で照合作業を行う場合に限らず、現存の照合作業済みのデータを確認する意味でも有用であると思われる。

また、実際の照合作業においては、人手判定の担当者の人数、それに関わる経費および時間も検討する必要があり、人手判定の対象とできるレコードペア数も限られる。自動判定を積極的に導入することで、人手判定の対象を減らすことができるが、その誤判別を把握および制御することが重要である。

図1. 同一人ペア確率と累積同一人ペア数。ファイルBのすべてのレコードに対して、ファイルAのすべてのレコードとの(9)式の同一人ペア確率を算出した。横軸に同一ペア確率 $10^2\alpha$ を、縦軸に区間 $I^\alpha = [\alpha, 1]$ に含まれる(10)式の同一人ペア数 $q^{(S)}(I^\alpha)$ を与えた。

まず、(9)式の同一人ペア確率が高いペアは同一人ペアとして自動判定すべきであろう。図1に、同一人ペア確率とその値までに含まれる累積同一人ペア数を与える。図から分かるように、同一人ペア確率が100%に近いところに同一人ペアは集中しており、以後なだらかに増えた後、0%の近くでも100程度の同一人ペアの存在がうかがえる。図2に、同一人ペア確率が高い方から自動判定で同一人ペアとしたときの累積別人ペア数を示す。例えば、男性において別人ペアなのに誤って同一人ペアとしてしまうペアを高々1ペアに抑えたければ、同一人ペア確率が100%から97.5%までのレコードペアを同一人ペアとして自動判定すればよい。図3には、同一人ペア確率が低い方から自動判定で別人ペアとしたときの累積同一人ペア数を示す。この場合も同様に、同一人ペアを1つ程度なら誤って別人ペアと自動判定してよければ同一人ペア確率が0%から0.00005%までを別人ペアとする。自動判定の対象とされなかったレコードペアが人手判定の対象となるが、表5に、男性のレコードについて、誤判別の設定とそのときに人手判定の対象となるレコードペア数の例をあげる。例えば、同一人ペア確率の高い方から累積別人ペア数が1になるまでは自動判定で「同一人ペア」、低い方から累積同一人ペア確率が1になるまでは自動判定で「別

図 2. 自動判定における誤判別制御のための累積別人ペア数. ファイル B のすべてのレコードに対して, ファイル A のすべてのレコードとの (9) 式の同一人ペア確率を算出した. 横軸に同一ペア確率 $10^2\alpha$ を, 縦軸に区間 $I^\alpha = [\alpha, 1]$ に含まれる (10) 式の別人ペア数 $q^{(D)}(I^\alpha)$ を与えた.

図 3. 自動判定における誤判別制御のための累積同一ペア数. ファイル B のすべてのレコードに対して, ファイル A のすべてのレコードとの (9) 式の同一人ペア確率を算出した. 横軸に同一ペア確率 $10^2\beta$ を, 縦軸に区間 $I_\beta = [0, \beta]$ に含まれる (10) 式の同一人ペア数 $q^{(S)}(I_\beta)$ を与えた.

人ペア」, とした場合, 全レコードペア数, 249,053,418 ペアのうち, 残りの人手による照合作業の対象となるレコードペア数は 1,191,271 となる. 照合作業に関わる経費が潤沢な場合には, 誤判別の設定から人手判定の対象レコードペア数を算出することもできるが, 予算が限られていれば, 人手判定が可能なレコードペア数の上限を設定し, そのときの誤判別を把握することになる.

今後の改良としては, 独立な項目グループの探索の他にも, 一致型の拡張もしくは緩和が考えられる. 例えば, 新旧漢字による「浜田」と「濱田」の表記違い, 平仮名およびカタカナによる「あきこ」と「アキコ」の違い, などは, 人手では同一視されており, 個人同定には有用な情報となる. したがって, 蓄積された同一人ペアの例などから同一視すべきレコード値を項目別にデータベース化し一致型を求める際に参照することで, 経験則を反映したより効率的な同定処理が可能になると思われる.

謝 辞 本稿の改訂にあたり、査読者の方々から大変有益なご意見を頂きました。ここに感謝の意を表します。なお、この研究の一部は、厚生労働科学研究費補助金「第3次対がん総合戦略研究事業」および文部科学省学術振興会からの科学研究費補助金による援助を受けています。

参 考 文 献

- D'Andrea, N.S. and Bois, D.J. (1969): A solution to the problem of linking multivariate documents, *Journal of the American Statistical Association* **64**, 163–174.
- Dunn, H.L. (1946): Record Linkage, *American Journal of Public Health* **36**, 1412–1416.
- Fellegi, I.P. and Sunter, A.B. (1969): A theory for record linkage, *Journal of the American Statistical Association* **64**, 1183–1210.
- Jaro, M.A. (1995): Probabilistic linkage of large public health data files, *Statistics in Medicine* **14**, 491–498.
- 近藤久義, 中村 剛, 三根真理子, 森 弘行 (1979): 診療記録データベース作成に関する基礎的研究, 1, 電算機による重複チェックの効果的方法, *日本公衆衛生雑誌* **27**, 201–204.
- Newcombe, H.B., Kennedy, J.M., Axford, S.J. and James, A.P. (1959): Automatic linkage of vital records, *Science* **130**, 954–959.
- Newcombe, H.B. and Kennedy, J.M. (1962): Record linkage, *Communication of the Association for Computing Machinery* **5**, 563–566.
- Newman, T.B. and Brown, A.N. (1997): Use of commercial record linkage software and vital statistics to identify patient deaths, *Journal of the American Medical Informatics Association* **4**, 233–237.
- Ohshima, A., Sakagami, F., Hanai, A. and Fujimoto, I. (1979): A method of record linkage, *Environmental Health Perspectives* **32**, 221–230.
- 小野雅司, 柳井晴夫, 中江公裕, 豊川裕之 (1976): 疫学調査における重複のチェック方法について, *日本公衆衛生雑誌* **23**, 9–16.
- 大瀧 慈, 務中昌己, 栗原登, 早川式彦, 山本修, 上岡洋史, 隅田治行, 平岡正行 (1982): 被爆者人口構築における関連資料間のレコードリンケージのための照合理論について, *広島医学* **35**, 3, 86–88.
- Ohtaki, M., Hayakawa, N., Kurihara, N. and Munaka, M. (1985): A mathematical theory of identification in record linkage, *Statistical Methods in Cancer Epidemiology*, Radiation Effect Research Foundation.
- Tepping, B.J. (1968): A model for optimum linkage of records, *Journal of the American Statistical Association* **63**, 1321–1332.
- 柳川 洋, 種村道彦, 重松逸造 (1971): レコードリンケージに関する基礎的研究, (その1) 疾病届出情報における同一人重複のチェック方法について, *日本公衆衛生雑誌* **18**, 487–493.

(2005 年 3 月 24 日受付 5 月 16 日最終修正 5 月 25 日採択)

著者連絡先: 大瀧 慈

〒734-8551 広島市南区霞 1-2-3 総合研究棟 4F 計量生物

A Statistical Method for Automatic Identification in Record Linkage

**Kenichi Satoh¹, Norihiko Hayakawa², Haruyuki Sumida³
and Megu Ohtaki^{1,*}**

¹ Department of Environmetrics and Biometrics,

² Department of Epidemiology,

³ Radiation Research Center for Frontier Science Facilities for Radiation Experiments,
Research Institute for Radiation Biology and Medicine, Hiroshima University, Kasumi 1–2–3,
Minami-ku, Hiroshima 734–8551, Japan

Abstract

As computing environments improve, many data files can be created and stored easily. Some of them might be related to each other and include records on the same person. For example, hospitals have clinical records and the government has vital statistics. We often need to link these records together in medical, epidemiological or sociological research. In Japan, however, no identification number is available like the Social Security Number in the USA. Therefore, we have to link records by using common key fields (or variables) among those files, such as family name, given name, address, sex, birthdate, etc. The procedure is known as record linkage and many empirical approaches have been considered. We discuss a mathematical model based on a pair of binomial distributions and propose a statistical method for automatic identification using frequencies of fixed record values.

Key words: identification criterion, mixed binomial distribution, Poisson distribution

*Corresponding author

(Megu Ohtaki)

Received March 24, 2005; Received in final form May 16, 2005; Accepted May 25, 2005.