の効果によって死亡率が低下しているリンパ性白血病，さらにこれら以外の悪性新生物を比較対照として，出生コホート別の死亡率の変化を検討した．

対象と方法

北海道の人口動態統計（1979年～2001年）の資料中の「副腎の悪性新生物」（ICD9では1940，ICD10ではC74），「リンパ性白血病」（LL；それぞれ204，C91），およびこれら以外の悪性新生物（その他の悪性新生物）の1～9歳における出生10万対の2001年末現在の累積死亡率を，

A群：1979～1983年出生

B群：1984～1987年出生

C群：1988～1991年出生

の3群について計算した．この区分は1984年が札幌市でのHPLC開始年，1988年が札幌市以外の北海道でのHPLC開始年であることによる．

転入・転出については考慮しなかった．北海道は全都道府県中，転入・転出率が最も低い[5]．また，北海道内で発生した悪性腫瘍の児はほとんど全員が道内の病院を受診しており，道外で発生した悪性腫瘍の児が北海道内の病院を受診することはほとんどない[6]．

人口動態統計の「死亡」においては，部位別分類が用いられているため，NBという項目はない．しかし，小児においては，「副腎の悪性新生物」は事実上すべてNBであるし，少なくとも人口動態統計の数字の上では，NBによる死亡の大部分は「副腎の悪性新生物」によるものである．厚生省の研究班であった塙・月本班は，すべての部位のNB死亡を検討し，その素データを公表した．これと，人口動態統計「副腎の悪性新生物」による死亡と突き合わせると，死亡の大部分（91.6%）は副腎原発によるものであった[7),8)]．また，厚生労働省の研究班である林班の解析結果によれば，1995年から2000年に出生して2001年までに死亡した神経芽腫106例のうち，C74で示される副腎の悪性新生物は99例（93.4%）であった[4)]．また，北海道小児悪性腫瘍登録の資料によれば，今回対象とした期間に出生し，かつ10歳未満で診断された副腎癌の例はなかった．

結　果

表1にHPLC受検率と死亡率を示す．A群のHPLC受検率は0%で，この群の死亡率を100とした場合の数字も併記した．B群のHPLC受検率は約24%である（札幌市のみの時代）が，C群は全道の児がHPLCでスクリーニングされたため約85%と高くなった．LLはA群からB群にかけての低下が著しいが，B群からC群にかけてはかえって上昇していた．これに対し，NBはB群からC群にかけての低下が大きく，LL以上の減少を示しており，これは先に報告した全国における傾向とほぼ同じである[9)]．「その他の悪性腫瘍」の死亡率はほとんど変化していない．

治療もMSも寄与がないとすればB群もC群も死亡率はA群と同じになるとし，死亡数の減少に対する治療とMSの寄与をそれぞれX，Yとすると，MS非受検者はXのみの寄与を受け，受検者はXとY両方の寄与を受ける．

表1．北海道における出生群別のHPLC受検率（括弧内は分母が全出生数，分子がHPLC受検数）および神経芽腫・リンパ性白血病・その他の悪性新生物による1歳以上10歳未満の出生10万対死亡率（括弧内左側は実数；右側はA群を100とした場合の%）．

群	出生年	HPLC受検率		死亡率		
				神経芽腫	リンパ性白血病	その他の悪性新生物
A	1979-83	0%	(0/371707)	5.92 (22; 100%)	9.42 (35; 100%)	22.06 (82; 100%)
B	1984-87	24.2%	(63343/261789)	5.73 (15; 96.8%)	6.88 (18; 73.0%)	19.86 (52; 90.0%)
C	1988-91	84.5%	(188362/222799)	4.04 (9; 68.3%)	7.18 (16; 76.3%)	21.10 (47; 95.6%)

B群においては

非受検者の死亡期待数は

198446×5.92＝11.74人，

受検者の死亡期待数は

63343×5.92＝ 3.75人，

C群においては

非受検者の死亡期待数は

34437×5.92＝ 2.04人，

受検者の死亡期待数は

188362×5.92＝11.15人

であるから，以下の連立方程式ができる．

B群　11.74X＋3.75 (X＋Y) ＝0.494

（ここに 2.61789×5.92－15＝0.494）

C群　2.04X＋11.15(X＋Y)＝4.187

（ここに 2.22799×5.92－9＝4.187）

これを解くと，X＝－0.08，Y＝0.47 が得られる．

考　察

今回の検討では，北海道において 1979 年 1 月 1 日から 1992 年 12 月 31 日までに出生した児の中で，2001 年 12 月 31 日までに 1 歳以上 10 歳未満で死亡した例について分析したことになる．つまりその性格はコホート研究である．また，定性法と HPLC とでは症例発見能力において大差があり，両者は区別して扱う必要があるため，今回は HPLC のみに着目したのである．つまり，定性法の死亡率減少に対する寄与はまったくないものと仮定して計算を行ったのである．今回，「治療」としたのは，厳密には「HPLC 以外の要因」である．実際には，札幌市以外の北海道において実施されていた DIP 法が B 群において多少の寄与をしているであろうから（A 群においては札幌市の TLC がやはり死亡率を多少は下げているであろうが，A 群の死亡率は連立方程式の両方の式に使用されているため，TLC の寄与分は結果には影響しない），その分が今回の「治療」の寄与から差し引かれるため，「治療」の寄与は今回の結果よりもう少し小さいことになる．

小児の悪性新生物の中で，治療の進歩の著しいものとして，急性リンパ性白血病（ALL）があげられる．人口動態統計の「死亡」では 204 (ICD9)と C91(ICD10)は「リンパ性白血病」であるが，小児では慢性リンパ性白血病は事実上ありえないため，これらのデータは即ち ALL と考えることができる．NB が ALL 以上の死亡率低下を示したということは，もし NB 死亡率の減少が専ら治療の進歩によるものであれば，ALL の治療を上回る進歩があった，ということになるであろう．それが具体的に如何なる治療法によるのかは，NB 死亡率の減少は専ら治療の進歩によるものであると主張する研究者から，未だかつて示されたことはない．

その他の悪性新生物の死亡率の減少傾向は明らかではなかった．全国的に見た場合，たとえば軟部悪性腫瘍の死亡率はここ 20～30 年間ほとんど変化しておらず[10)-12)]，今回の北海道のデータも，これと矛盾しないものである．軟部悪性腫瘍の死亡率は 5 歳未満では減少しているが，5 歳以上では増加しており，結局 15 歳未満全体としてはほとんど変化していない．これは，治療の進歩によって延命がなされ，死亡年齢が多少上昇したことによる変化である[10)]．MS を実施していないフランスにおいて，これと同様な変化が NB 死亡率にも起こっていたことは既に報告した[9)]．

治療の寄与が，今回の計算ではマイナスとなった．このことは治療をすると死亡率が増えることを意味する．これは，化学療法の副作用や手術の合併症などによって，かえって寿命が短縮される場合があるからかも知れないが，全体として考えた場合，計算誤差の範囲と見なす方が自然であろう．今回の結果の意義は，プラスやマイナスの符号ではなく，治療の寄与より NBMS の効果が遥かに大であるということにある．実際，今回の結果は，以前，全国の人口動態統計を使用した計算[7), 8)]とほぼ同様の結果であって，MS 受検者の死亡の約 47%を減少させたことを意味し，林班が出した結果[2)-4)]とも矛盾しないものである．

NBMS に対する反対論は，初めに結論があ

り，その時々の事情に応じてそれを正当化する形でなされてきた経緯がある．1990 年前後にはその論拠は「死亡率が減少していないからNBMS は無意味」であった．ところが，HPLC が普及してその効果で死亡率が減少してくると，「死亡率の減少が有意でないから」に変化した．1990 年代末から 2000 年代初めにかけて，久繁班が後ろ向きコホート研究で有意な死亡率の減少を報告すると「北米[13]やドイツ[14]で否定的な結論が出たから」に変わり，林班がさらに前向きコホートで同様の有意な死亡率の減少を報告した今では「過剰治療が 1 人にでもなされるような検診は，どれだけ死亡率を減少させようと実施されるべきではない」となっている．この論法で行くと，たとえば予防接種さえも，その副作用から場合によっては死亡する場合もあるわけだから，実施することができなくなるであろう．少子化がますます進行する現在，とにかく 1 人でも多くの児の命を救わなければならないのに，この意見には将来の子孫を救うという観点が全く欠落している．これから成人して次の世代を生む児 1 人の生命を救うことは，既に子孫を作り終わった成人 1 人の命を救うこととは，同じ 1 人の命を救うことであっても，その意味は全く異なる．今の 1 人の児の命が失われると，将来その児から生まれるはずだった子孫の命まで失われてしまう．逆に，1 人の児を救うことは，将来の何人もの子孫を救うという意味があるのである．

文　献

1) 久繁哲徳：神経芽細胞腫スクリーニングの評価．厚生科学研究費補助金（子ども家庭総合研究事業）総括研究報告書　p167-174, 1999.
2) 林　邦彦，藤田利治，片野田耕太　他：全国乳児コホートを対象とした神経芽細胞腫死亡における受検・未受検の比較研究．厚生労働科学研究費補助金　難治性疾患克服研究事業　マススクリーニングの効率的実施及び開発に関する研究．平成 15 年度総括・分担研究報告書　2004, 122-130.
3) Hayashi K, Fujita T, Katanoda K et al.: Effectiveness of mass-screening program on neuroblastoma mortality in 1995-2000 birth cohort of Japan: Nationwide Neuroblastoma Mortality Study. 11th Conference of Advances in Neuroblastoma Research. Genova, Italy, June 16-19, 2004.
4) 林　邦彦，藤田利治，片野田耕太　他：全国乳児を対象とした神経芽細胞腫死亡における受検・未受検の比較研究．第 32 回日本マス・スクリーニング学会特別講演．2004 年 10 月 8 日，仙台市．
5) 都道府県別にみた自然増加率および社会増加率．1990 人口の動向，人口問題研究会編，厚生統計協会，1991, p159.
6) 西　基：神経芽腫のマス・スクリーニング．日マススク誌　10: 5-16, 2000.
7) 西　基，武田武夫，畑江芳郎　他：神経芽腫死亡の減少に対する HPLC マス・スクリーニングの効果．日マススク誌　11: 45-49, 2001.
8) Nishi M, Takeda T, Hatae Y et al.: Contribution of HPLC mass screening for neuroblastoma to a decrease in mortality. J Exp Clin Cancer Res 21: 73-78, 2002.
9) 西　基，武田武夫：マススクリーニングが神経芽腫の治療成績に与えたインパクト．小児外科　36: 63-69, 2004.
10) Nishi M, Hatae Y: Epidemiology of malignant neoplasms in soft tissue during childhood. J Exp Clin Cancer Res 23: 437-440, 2004.
11) 畑江芳郎，西　基：悪性軟部腫瘍の疫学と臨床像．小児外科　34: 362-370, 2002.
12) Nishi M. Epidemiology of childhood malignancies in Japan. Recent Res Devel Cancer 6: 81-88, 2004.
13) Woods WG, Tuchman M, Robison LL et al.: A population-based study of the usefulness of screening for neuroblastoma. Lancet 348: 1682-1687, 1996.
14) Schilling FH, Spix C, Berthold F et al.: Neuroblsatoma screening at one year of age. N

Engl J Med 346; 1047-1053, 2002.

受付日：平成17年 2 月 14 日
受理日：平成17年 9 月 20 日

Mortality of neuroblastoma in birth cohorts in Hokkaido Prefecture

Motoi NISHI[1)], Susumu IIZUKA[2)], Junji HANAI[3)], Kozo FUJUTA[3)], Hisae ICHIMIYA[4)], Toshiyasu TANAKA[4)]

1) Department of Fundamental Health Sciences, Health Sciences University of Hokkaido
2) Hokkaido Cancer Center
3) Sapporo City Institute of Public Health
4) Hokkaido Pharmaceutical Association Public Health Examination Center

応用統計学 Vol. 34, No. 1 (2005), 1–13

研究論文

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際的，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，$\mathcal{A} = \{\boldsymbol{a}_i | i = 1, \cdots, N_A\}$, および $\mathcal{B} = \{\boldsymbol{b}_i | i = 1, \cdots, N_B\}$, と表現できる．ただし，$\boldsymbol{a}_i = (a_{i1}, \cdots, a_{iK})$, $\boldsymbol{b}_i = (b_{i1}, \cdots, b_{iK})$. 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 $(\boldsymbol{a}, \boldsymbol{b})$ 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，

$$\boldsymbol{t}(\boldsymbol{a}, \boldsymbol{b}) = (t_1, \cdots, t_K), \tag{1}$$

を得る．ただし，$k = 1, \cdots, K$ 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
$\boldsymbol{a}$	佐藤	健一	明治	4	4	9
$\boldsymbol{b}$	佐藤	春夫	明治	4	4	9
$\boldsymbol{t}(\boldsymbol{a}, \boldsymbol{b})$	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．

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

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，

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

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，$y^{(j)}(t) \sim Poisson(\lambda^{(j)}(t)), j \in \{S,D\}$, および，$y(t) \sim Poisson(\lambda(t))$, とする．ただし，$Poisson(\lambda)$ 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，

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

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，$\{t_m | m = 1, \cdots, 2^K\}$ 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．

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

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，

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

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．

$$p^{(j)}(\boldsymbol{t}) = \prod_{l=1}^{L^{(j)}} p_l^{(j)}(\boldsymbol{t}_l^{(j)}), \qquad \boldsymbol{t} = (\boldsymbol{t}_1^{(j)}, \cdots, \boldsymbol{t}_{L^{(j)}}^{(j)}), \qquad j \in \{S, D\}, \tag{7}$$

ただし，$\boldsymbol{t}_l^{(j)}$ 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 $\boldsymbol{t}_l$ 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 $\boldsymbol{t}$ 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．

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

Fellegi and Sunter(1969)および Tepping(1968)では，オッズ比型，$r^{(S)}(\boldsymbol{t})/\{1 - r^{(S)}(\boldsymbol{t})\}$ 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 $\boldsymbol{b}$ 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 $\boldsymbol{b}$ がデータファイル A の別人に関するレコードと一致型 $\boldsymbol{t}$ を持つ確率を $p^{(D)}(\boldsymbol{t}|\boldsymbol{b})$ とすれば，ファイル B のレコード $\boldsymbol{b}$ 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，

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

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 $r_1^{(S)}, \cdots, r_{N_A N_B}^{(S)}$ 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，$J(I) = \{i | r_i^{(S)} \in I, i = 1, \cdots, N_A N_B, \}$ 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，

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

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．

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

さらに，任意の一致型 $\boldsymbol{t}_0$ を持つレコードペアの集合に対して，

$$\frac{r^{(S)}(\boldsymbol{t}_0)}{\sum_{(a,b) \in C(\boldsymbol{t}_0)} r^{(S)}(\boldsymbol{t}(\boldsymbol{a}, \boldsymbol{b}) | \boldsymbol{b})} = 1,$$

を考えることもできる．ただし，$C(\boldsymbol{t}_0) = \{(\boldsymbol{a}, \boldsymbol{b}) \in \mathcal{A} \times \mathcal{B} | \boldsymbol{t}(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{t}_0\}$ 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，

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

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 $\boldsymbol{t}$ 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，$\mu_k^{(D)}, k = 1, \cdots, K$ を既知パラメータとし，M 個のペア数データ $\{y(\boldsymbol{t}_m) | m = 1, \cdots, M\}$ から，未知パラメータ ξ および $\mu_k^{(S)}, k = 1, \cdots, K$ 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 $\hat{p}^{(j)}(\boldsymbol{t}), j \in \{S, D\}$ を用いて，$\hat{\lambda}^{(j)}(\boldsymbol{t}) = \hat{N}^{(j)}\hat{p}^{(j)}(\boldsymbol{t}), j \in \{S, D\}, \hat{N}^{(S)} = \min(N_A, N_B)\hat{\xi}, \hat{N}^{(D)} = N_A N_B - \hat{N}^{(S)}$, のように推定でき，(8) の同一人ペア確率が推定できる．

同様にして，レコード $\boldsymbol{b}$ を固定したときの同一人ペア確率の推定については，$\mathcal{A} \times \{\boldsymbol{b}\}$ 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 $f_k^{(D)}(\boldsymbol{b})$ とすると，同一人ペアの存在を無視することで $\mu_k^{(D)}(\boldsymbol{b})$ は $f_k^{(D)}(\boldsymbol{b})/N_A$ によって推定できる．このとき，定義から，

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

が成り立つことに注意する．なお，$p^{(D)}(\boldsymbol{t}|\boldsymbol{b})$ 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 $\boldsymbol{t}$ 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く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 $f_k^{(D)}, k = 1, \cdots, K$ 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 $\hat{\mu}_k^{(j)}, k = 1, \cdots, K, j \in \{S, D\}$ 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 $\hat{r}^{(S)}$, 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 $\boldsymbol{b}$ についてレコードリンケージを行うために，ファイル A におけるレコード値の頻度 $f_k^{(D)}(\boldsymbol{b}), k = 1, \cdots, K$ を調べ(表 4 参照)，$\hat{\mu}_k^{(D)}(\boldsymbol{b}) = f_k^{(D)}(\boldsymbol{b})/N_A$, $k = 1, \cdots, K$ とし，3) で推定された $\hat{\mu}_k^{(S)}, k = 1, \cdots, K$ 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，満たされていなければ，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 $m = 1, \cdots, 7$ までのデータだけを使用した．

	一致型						ペア数		同一人ペア確率		同一人ペア数	
m	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)$10^2 \hat{r}^{(S)}$, 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 $m = 1, \cdots, 4, 6, 7$ および $m = 11$ のデータだけを使用した．

	一致型						ペア数		同一人ペア確率		同一人ペア数	
m	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	1,163	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)$10^2 \hat{r}^{(S)}$, 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
$\boldsymbol{b}$	佐藤	春夫	明治	4	4	9
$N_A/f_k^{(D)}(\boldsymbol{b})$	307.99	969.19	2.84	18.77	13.21	42.9
$\boldsymbol{b}$	谷崎	潤一郎	明治	19	7	24
$N_A/f_k^{(D)}(\boldsymbol{b})$	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 $I^{\alpha}_{\beta} = [\beta, \alpha]$ に含まれるレコードペア数を算出した．

	別人ペア候補に含まれる下側累積同一人ペア数			
同一人ペア候補に含まれる上側累積別人ペア数	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 $100 \times \hat{r}^{(S)}$ 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．

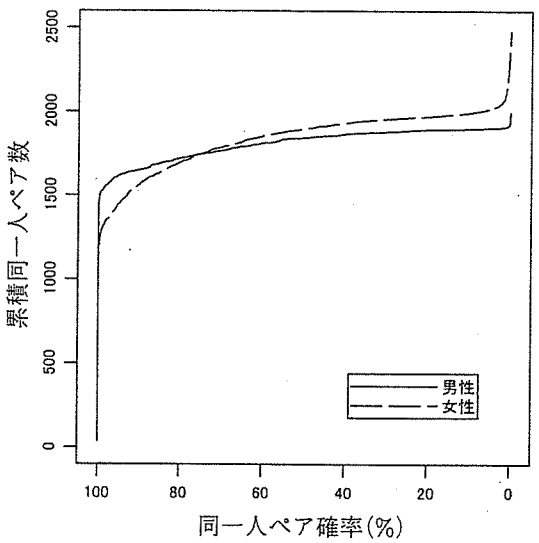

図 1. 同一人ペア確率と累積同一人ペア数．ファイル B のすべてのレコードに対して，ファイル A のすべてのレコードとの (9) 式の同一人ペア確率を算出した．横軸に同一ペア確率 $10^2\alpha$ を，縦軸に区間 $I^\alpha = [\alpha, 1]$ に含まれる (10) 式の同一人ペア数 $q^{(S)}(I^\alpha)$ を与えた．

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

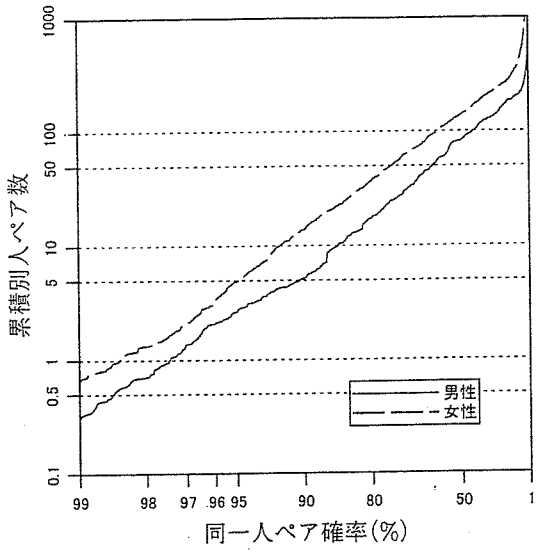

図 2. 自動判定における誤判別制御のための累積別人ペア数．ファイル B のすべてのレコードに対して，ファイル A のすべてのレコードとの (9) 式の同一人ペア確率を算出した．横軸に同一ペア確率 $10^2\alpha$ を，縦軸に区間 $I^\alpha = [\alpha, 1]$ に含まれる (10) 式の別人ペア数 $q^{(D)}(I^\alpha)$ を与えた．

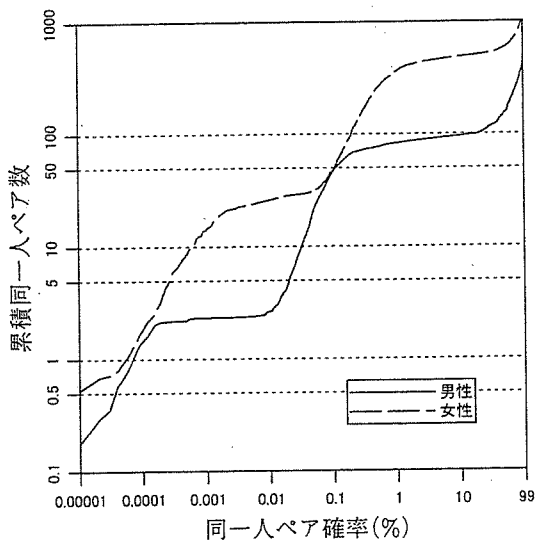

図 3. 自動判定における誤判別制御のための累積同一ペア数．ファイル B のすべてのレコードに対して，ファイル A のすべてのレコードとの (9) 式の同一人ペア確率を算出した．横軸に同一ペア確率 $10^2\beta$ を，縦軸に区間 $I_\beta = [0, \beta]$ に含まれる (10) 式の同一人ペア数 $q^{(S)}(I_\beta)$ を与えた．

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
TEL:.082–257–5852
E-mail: ohtaki@hiroshima-u.ac.jp

Japanese J. Appl. Statist. 34 (1) (2005), 1–13

A Statistical Method for Automatic Identification in Record Linkage

Kenichi Satoh[1], Norihiko Hayakawa[2], Haruyuki Sumida[3] 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
E-mail address: ohtaki@hiroshima-u.ac.jp (Megu Ohtaki)

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

IVR会誌 Jpn J Intervent Radiol 20:376-384, 2005.

2. 統計解析で何が得られるか，その可能性と危険性

広島大学原爆放射線医科学研究所，放射線システム部門計量生物研究分野
大瀧 慈

Department of Environmetrics and Biometrics, Research Institute for Radiation Biology and Medicine, Hiroshima University
Megu Ohtaki

Abstract

What Provide by Statistical Analyses, Its Possibility and Danger

With the recent rapid improvement of performance of computer and network environment, various sophisticated statistical theories that were "castles in the air" can now be applied to data analysis easily. In view of such a background, the purpose of this paper is to outline several useful statistical methods such as summarization of data using classical descriptive statistics, correlation analysis, regression analysis and the latest methods called "computer intensive techniques". It is also discussed how to apply them and how to interpret the results with possible pitfalls that should be avoided.

はじめに

最近のコンピュータやネットワークの性能の飛躍的向上に伴って，各種の大容量のデータベースが構築されるとともに従来は「絵に描いた餅」であった高度で複雑な統計理論が，現実に手軽にデータ解析へ適用できる環境になってきている。これまでに提案されている統計的手法には，どのようなものがあるのか，また，如何に適用できるのか，その際に注すべきその落とし穴はないのか，という観点に立って，統計的データ解析の現場でよく使われている（または，その可能性の高い）古典的な記述統計および相関分析と回帰分析をはじめ，コンピュータ集中技法と称される最新の各種技法について論ずる。なお，本論文で紹介している各統計解析は，全て当研究室で開発した独自のプログラム（インターネット上で公開中）を用いて行ったものである。

データの要約

1. 算術平均

統計解析の基本はデータの要約である。計量値データを対象にする場合にその最も古典的な手法が算術平均を求めること である。算術平均は，計量データ$y_1, \cdots, y_n$が与えられたとき，$\bar{y}=\frac{1}{n}\sum_{i=1}^{n} y_i$で定められるものであるが，その要約基準としての妥当性は，

$$S(\theta)=\sum_{i=1}^{n}(y_i-\theta)^2=n(\theta-\bar{y})^2+\sum_{i=1}^{n}(y_i-\bar{y})^2\geq S(\bar{y})$$

による。いわゆる最小自乗規準に基づくものである。多次元の場合も，データ$\boldsymbol{y}_i=(y_{i1}, \cdots, y_{ip})^{\mathrm{T}}, i=1, \cdots, n$の算術平均ベクトル$\bar{y}=\frac{1}{n}\sum_{i=1}^{n} y_i$に関して，

$$S(\theta)=\sum_{i=1}^{n}\|y_i-\theta\|^2=\sum_{i=1}^{n}(y_i-\theta)^T(y_i-\theta)$$

$$=n\|\theta-\bar{y}\|^2+\sum_{i=1}^{n}\|y_i-\bar{y}\|^2\geq S(\bar{y})$$

が満たされているので，1次元の場合と同様な意味付けが可能である。ここで，$\|z\|$はベクトルzのノルムである。

平均値は計算が容易であり，上記のような分かり易い特性を持っている。その一方，極端に大きい（または小さい）値（このような値は"外れ値"と呼ばれている）がデータに混入している場合には，それらの影響を強く受けてしまうという欠点も持っている。表1は5個の標本から成る3組の人工データを示す。これらのデータはいずれも平均値が10.0である。データAは，平均値が要約統計量として扱われても問題が無い思われる場合であり，データBとCは，何れも問題があると思われる場合を示している。データBの場合は，右裾が重たい非対称な分布をしており，データCには1個の外れ値が存在している。

算術平均による要約では，データが1個の平均的な値（多次元の場合は1個のベクトル）の周りにほぼ対称的に分布していることが暗黙の前提とされている。もち

Key words
- *Outline of statistical methods*
- *Pitfall of analysis*
- *Statistical data analysis*

ろん，代数的には，算術平均を求める際にそのような前提は要求されていないが，算術平均値が与えられれば，一般的に人は「データはその値の周りに対称的に分布している」ということをイメージしてしまう傾向が強く，誤解の元となる危険がある。表1のデータBやデータCの場合のように1個の代表値による要約に無理が有る場合に対して，対数変換などの変数変換による分布の対称化処理や，複数個の代表値による要約が有効である。

2. 主要点解析

主要点解析（Principal Points Analysis）は，*k*-means法とも呼ばれている要約手法で，複数個の代表値による要約を実現する手法である[1,2]。この解析では，*n*個のR^p内のデータ$y_1, \cdots, y_n$が与えられたとき，

$$\sum_{i=1}^{n} \min_{j=1,\cdots,k} \| y_i - a_j \|^2$$

の値を最小とするR^p内の*k*個の局所平均ベクトル$a_1, \cdots, a_k$を求め，要約とする。この解析手法は，郵便ポストの最適な配置の問題や天気図の分類などへの応用が試みられている[3]。

表1 同じ算術平均値を持つ3個のデータセット

データ識別	データA	データB	データC
1	5.00	2.00	1.00
2	8.00	5.00	2.50
3	10.00	9.00	2.50
4	12.00	14.00	4.00
5	15.00	20.00	40.00
算術平均値	10.00	10.00	10.00
偏差平方和	58.00	206.00	1129.50
分散	14.50	51.50	282.38
標準偏差	3.81	7.18	16.80
中央値	10.00	9.00	2.50

Box-Cox変換

正値をとる変数*y*に対して，下記のようなBox-Cox変換を適用し，変数の分布の非対称性や歪みを緩和 させることがある[4]。

$$y \to y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} + \lambda, & \lambda \neq 0, \\ \log y, & \lambda \neq 0, \end{cases} \qquad (1)$$

この変換族は，並行移動変換を無視すると，対数変換（$\lambda=0$），恒等変換（$\lambda=1$），平方根変換（$\lambda=1/2$），立方根変換（$\lambda=1/3$）などを特例として含んでいる。なお，Cox-Box変換のオリジナル版では，$\lambda \neq 0$の場合，

$$y \to y^{(\lambda)} = \frac{y^\lambda - 1}{\lambda}$$

となっているが，$\lambda=1$の場合に恒等変換にならない不都合が生じるため，(1) 式では変形を行っている。この"Modified" Box-Cox変換がオリジナル版の変換と同じ統計的特性を持つことについては，濱崎らによる研究がある[5]。

相関分析

2個の変数の間に相関があるか否かを調べるために，視覚的方法として相関プロット（散布図），数値的方法として相関係数が使用されることが多い。2変量間の（積率）相関係数とは，データ$\{(x_i, y_i), i=1, \cdots, n\}$与えられているとき，

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (2)$$

で定められるものである。以下，これらの方法で，相関の強さを把握する場合に気をつけなければならない点について，実例を挙げながらいくつか紹介する。

図1aは，某大学医学部の学生の2年生における身長（cm）と体重（kg）のデータの散布図を示す。このデータに対して算出される身長と体重の相関係数は，0.76で，決定係数は$0.76^2=0.58$である。これらの結果から，「体重は身長により5割以上の規定されている」と解釈してよいであろうか？その答えは，「はい」というより「いいえ」に近いかもしれないのである。実は，このデータには男子学生と女子学生に関するものが，それぞれ，130名分と40名分含まれている。各性別の群毎にマークを変えてプロットした散布図を**図1b**に示す。また，群別の身長と体重の相関係数は，男子学生で0.54，女子学生で，0.52であり，それに基づく決定係数は，何れも28％前後となる。即ち，最初の解析で得られた高い相関係数は，性別を介してえられたものであり，その効果を除去すると，体重に対する身長の寄与は3割以下となりかなり小さくなってしまうことが分かる。一般に母集団が複数の群の合併により構成されている場合には，それぞれの群における相関係数と合併された群全体におけるそれが一致するとは限らない。解析の目的によっては調整を行う必要が出てくるが，その調整は，データの中に群を識別する変数の値が存在する場合にのみ可能であり，調整された相関係数のことは偏相関係数と呼ばれている。

最近，遺伝子の発現状況を把握するためにマイクロアレイデータが作成され，解析が試みられるようになっている。素朴なマイクロアレイデータにはその作成過程における様々な技術的要因により無視できない偏りを内在させていることが多い。**図2**は，ある同一の試料に関して2回作成されたcDNAマイクロアレイについて，第1回目と第2回目に測定された遺伝子発現強度データの相関を示すものである。(a) がノーマリゼーション[6]の処理前，(b) がノーマリゼーション後のものである。相関係数は，それぞれ，0.42と0.38である。ノーマリゼーション処理により第1回目と2回目に共通したチィップの癖（偏り）が軽減され，その結果，相関が低くなっていることが分かる。いま，繰り返し観察における再現性をこの相関係数で評価できるかどうかという問題について考察してみよう。第1回目と第2回目の観察値，X_iとY_iは，それぞれ，

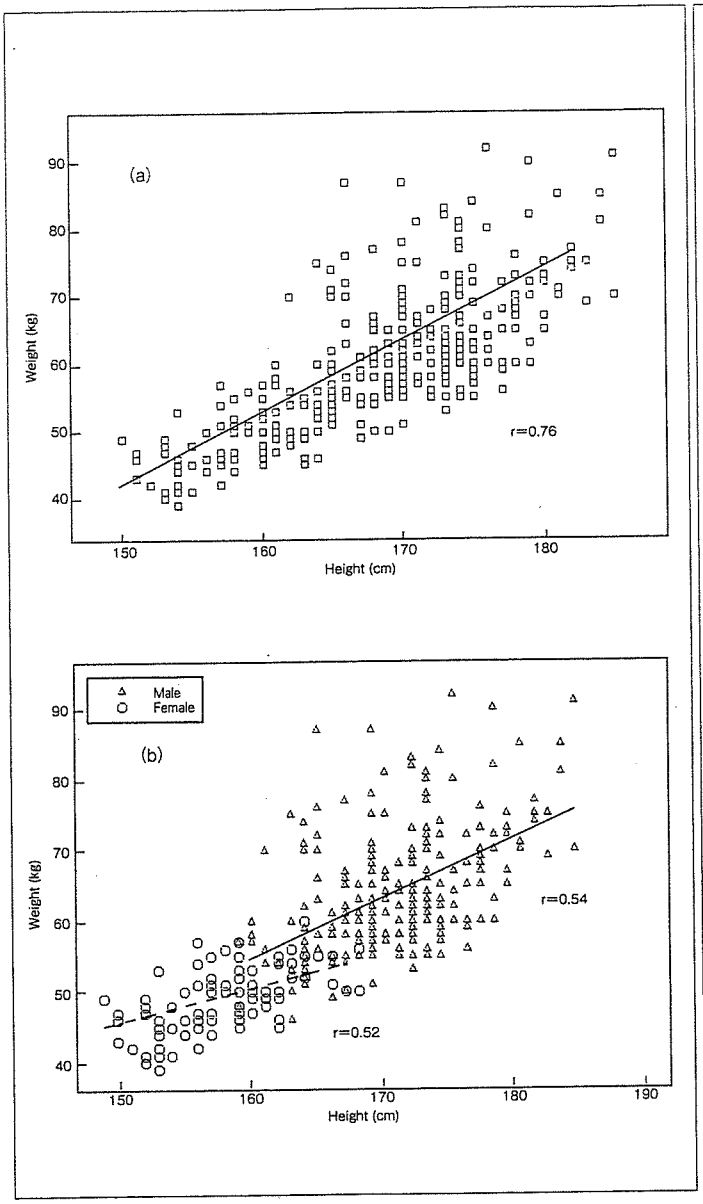

図1　某大学医学部2年生における身長と体重の散布図および最小自乗法により当てはめられた直線

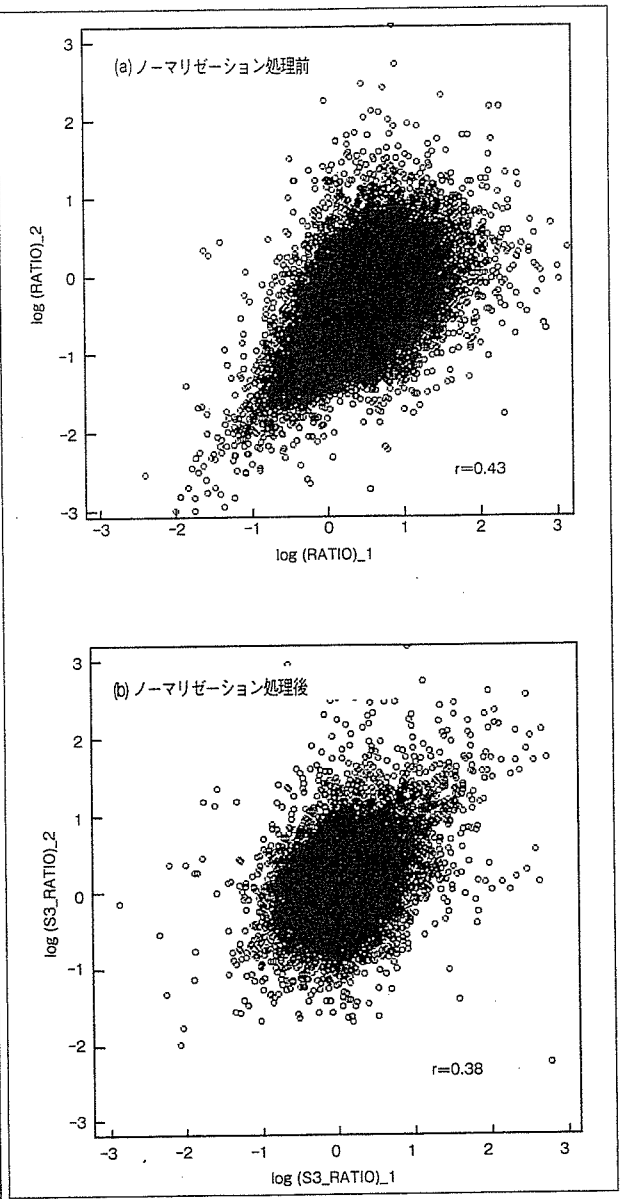

図2　繰り返して作成されたマイクロアレイデータにおける1回目と2回目の遺伝子発現強度の相関図，ノーマリゼーション処理前 (a) とノーマリゼーション処理後 (b)

$$X_i = \tau_i + \beta_i + \varepsilon_i,$$
$$Y_i = \tau_i + \beta_i + \delta_i,$$

により記述できるものとする。ここで，τ は真の信号，β は第1回目と第2回目に共通な偏り，ε と β はそれぞれ，第1回目と第2回目の誤差で，互いに独立であり，次のような分散を有しているものとする。
$Var(\tau_i) = \sigma_\tau^2, Var(\beta_i) = \sigma_\beta^2, Var(\varepsilon_i) = Var(\delta_i) = \sigma_\varepsilon^2$.
このとき，第1回目と2回目の観察値の相関係数に関して，

$$Var(X_i + Y_i) = 4\sigma_\tau^2 + 4\sigma_\beta^2 + 2\sigma_\varepsilon^2,$$
$$Var(X_i - Y_i) = 2\sigma_\varepsilon^2$$

であることに注目すると，(2) より下記の式が導かれる。

$$R \approx \frac{2\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\hat{\sigma}_\tau^2+\hat{\sigma}_\beta^2}{\hat{\sigma}_\tau^2+\hat{\sigma}_\beta^2+\hat{\sigma}_\varepsilon^2} \quad (2)$$

したがって，(3) 式の $\hat{\sigma}_\beta^2$ で示される共通の偏りの分散が大きい場合に，相関係数は 結果的に1に近づくことがわかる。

同様な機序により，各家庭で測定される夫婦の体重の測定値の間に，見かけの強い正の相関を生じさせる可能性がある。もしある家庭の体重計は初期状態で5kg を示す状態にあり，別の家庭の体重計の初期状態が－3kg であるというように，各家庭の体重計において十分な0点調整が行われていないならば，夫婦の体重間に共通の偏りが生じ，それにより夫婦の体重間に見かけの正の相関が現れうるからである。これらの例より，繰り返し観察データの相関解析を行う場合，「高い相関係数は再現性が高いことを必ずしも意味しないし，高い精度の結果を裏付けてもいない」，また，「低い相関係数は，必ずしも観測精度が低いことを意味しない」ということに留