計算言語学１０－１
（１９７７．６．２４）

同姓同名の発生頻度

田　中　康　仁　（日本ユニバック㈱）

1. はじめに

データ処理の中では時々おもしろいことがある。データ量が少ない時はあまり問題にならない事が，データの量が或る一定量を越えると急に大きな問題になる。
そのため従来の考え方や方法では解決出来ないような状態になる。
多量のデータを１ケ所に集中し統一された処理方式で処理することは経費や時間を削限し，大変効率良い処理方式である。しかしこれら合理的処理にはプラスの効果ばかりでなくマイナスの効果が生じてくる。例えば１００万件のデータについて名寄せ処理を行うとすると漢字の姓名だけでは名寄せは考えられず別の情報を附加しなければ個人の識別は行えない。そこでこの同姓同名はどの程度発生するか漢字の姓名と仮名の姓名について調べてみた。

2. 同姓同名の調査目的

同姓同名の調査を行なうにあたって考えた目的は次のものである。

(1) 同姓同名の発生率を定量的に分析する。
- データ量の増加にともない同姓同名の対象人員の増加度合
- 同姓同名の最も多く発生する件数

(2) 同姓同名にどの程度の情報を附加することによって個々人を識別することが可能であるか。

(3) 同姓同名を識別するための指導教育の資料を得る。

(4) 漢字と仮名では同姓同名の発生頻度は異ってくるか

以上のような目的で調査を行ってみました。

3. 同姓同名の調査（漢字）

漢字の姓名のファイルにより１０７万件のデータを分析した。このデータは同一人が二重に登録されることのないように十分チェックが行われたファイルである。このデータの対象地区は関東地方が中心になっているためデータには少しかたよりがあると思われる。
調査内容をまとめると次のようになる。

調査内容

(i) 調査年月日　：　昭和５１年４月
(ii) 調査対象件数　：　1,073,517件
(iii) 調査データ　：　漢字の姓名
(iv) データの性質　：　同一人のチェックが十分行われている。
(v) データの対象地区　：　関東地方

１０７万件のデータをランダム・ナンバーを使用し，１万件，５万件，２６万件，５３万件のデータに分割し分析を行った。

4. 調　査　結　果

まず1万件のデータによる調査結果をあげると表1のようになる。

1万件のデータによる同姓同名調査結果

同姓同名人数	件　　数	人　　数
5	1	5
3	3	9
2	76	152
1	10,570	10,570
計		10,736

表　1

この表の見方は次のように見て下さい。5名の同姓同名が1組あり，次に3名の同姓同名が3組あり，2名の同姓同名が76組あったことを表わしている。

次に5万件のデータによる分析結果をあげておく。表の見方は1万件の場合と同じである。

5万件のデータによる同姓同名調査結果

同姓同名人数	件　　数	人　　数
8	1	8
7	1	7
6	5	30
5	13	65
4	47	188
3	200	600
2	1,275	2,550
1	50,228	50,228
計		53,676

表　2

1万件のデータと5万件のデータを比べてみるとデータが増加するに従って急速に同姓同名の件数も増加していることが判る。1万件のデータでは同姓同名により漢字だけの姓名では識別のつかない人は166人で約1.54％ですが5万件のデータでは3,448人となり6.42％と急増している。

次に10万，26万，53万件のデータの詳細な分析結果は省略し，107万件のデータによる分析結果（表3）を示す。

１００万件のデータによる同姓同名調査結果

Seq	同姓同名人数	件　数	人　数
1	１８１	1	１８１
2	１３９	1	１３９
3	１２４	1	１２４
4	１０５	1	１０５
5	１０２	1	１０２
6	９６	1	９６
7	９３	1	９３
8	８９	1	８９
9	８５	1	８５
10	８３	1	８３
11	８１	1	８１
12	８０	1	８０
13	７５	1	７５
14	７３	2	１４６
15	７１	1	７１
16	６８	2	１３６
17	６７	2	１３４
18	６５	3	１９５
19	６３	2	１２６
20	６２	1	６２
21	６１	2	１２２
22	６０	2	１２０
23	５９	1	５９
24	５８	2	１１６
25	５７	3	１７１
26	５５	3	１６５
27	５４	2	１０８
28	５３	3	~~１０６~~ ７５９
29	５２	1	５２
30	５１	5	２５５
31	５０	2	１００
32	４９	6	~~２４５~~ ２９４
33	４８	8	３８４
34	４７	7	３２９
35	４６	１０	４６０
36	４５	4	１８０
37	４４	5	２２０
38	４３	9	３８７
39	４２	5	２１０
40	４１	1	４１
41	４０	9	３６０

Seq	同姓同名人数	件　数	人　数
42	３９	5	195
43	３８	16	608
44	３７	13	481
45	３６	13	468
46	３５	19	665
47	３４	16	544
48	３３	11	363
49	３２	21	672
50	３１	19	589
51	３０	18	540
52	２９	23	667
53	２８	23	644
54	２７	30	810
55	２６	34	884
56	２５	52	1,300
57	２４	52	1,248
58	２３	67	1,541
59	２２	69	1,518
60	２１	77	1,617
61	２０	83	1,660
62	１９	94	1,786
63	１８	113	2,034
64	１７	119	2,023
65	１６	162	2,592
66	１５	208	3,120
67	１４	250	3,500
68	１３	298	3,874
69	１２	372	4,464
70	１１	461	5,071
71	１０	666	6,660
72	9	813	7,317
73	8	1,180	9,440
74	7	1,733	12,131
75	6	2,517	15,102
76	5	4,447	22,235
77	4	8,433	33,732
78	3	19,203	57,609
79	2	71,137	142,274
80	1	714,964	714,964
	合　計		1,073,517

表　3

１０７万件のデータを調べると１８１人の同姓同名が発生している。この名前は鈴木和子という名前である。鈴木は関東地区に多い名前であるし，和子という名前は昭和以後の女性の名前としては大変多いものである。多い姓，名の結合が同姓同名を起しやすいという結果になっている。
１０７万人の中で同姓同名が起こるだろう人は３５万８千人になる。これは１０７万人のデータの３３％である。３人に１人は同姓同名になりうる相手があるということになる。これでは漢字の名前といえども個人を識別する能力は無くなっている。
各調査結果より同姓同名によって個人の識別ができない割合を調べてみると表４のようになる。

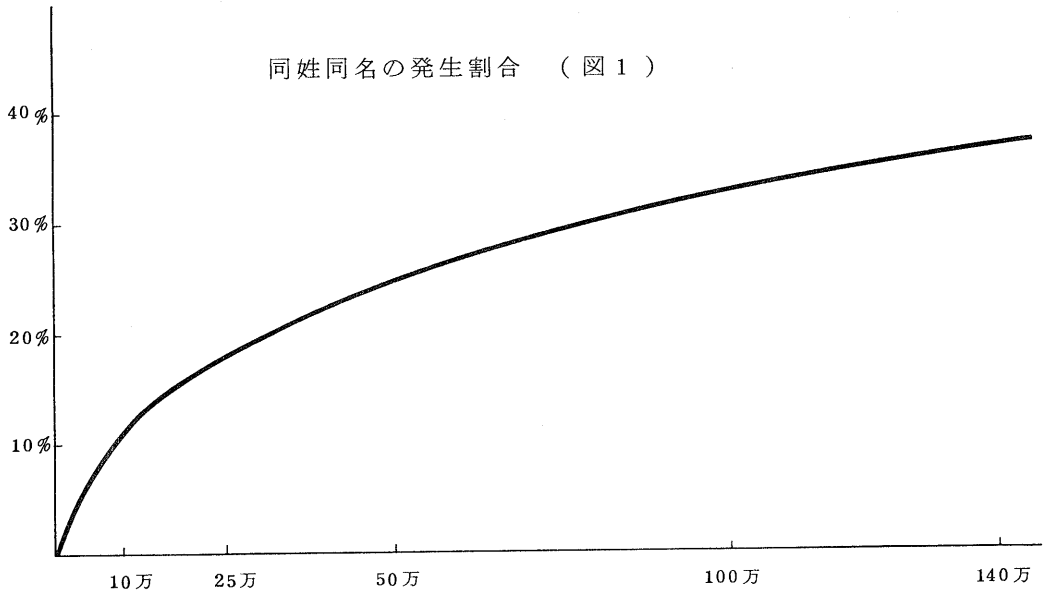
同姓同名の発生割合

サンプル No.	Data 件数	同姓同名の人数	%	同姓同名の最多人数
1	10,736	166	1.54	5人
2	53,676	3,448	6.42	8人
3	107,351	11,060	10.30	32人
4	268,377	47,195	17.58	36人
5	536,752	132,834	24.74	80人
6	1,073,517	358,306	33.38	181人

表　4

データの増加の割合に対して同姓同名の人数が急激に増えていることがわかる。しかし表４でもわかるように同姓同名の最も多く発生した人数はデータの増加にほぼ対応して増えている。
表４の結果を縦軸に同姓同名の発生割合（％），横軸に対象データ量を取りグラフを描くと図１のようになる。

同姓同名の発生割合　（図１）



この図は $y = a\sqrt{x}$ に似た曲線になっている。この曲線を手作業で近似するとデータ量が800万～1,000万件になると同姓同名の割合が100％近くなる。
漢字の姓名だけによって個人の識別は不可能になる。
同姓同名の起きやすい名前を男女別にあげると次のようになる。

男姓名で同姓同名の多いもの

1.	鈴木　実	9.	高橋　清
2.	田中　実	10.	佐藤　進
3.	鈴木　茂	11.	加藤　清
4.	鈴木三郎	12.	鈴木　博
5.	鈴木　清	13.	小林　茂
6.	~~鈴木~~斉藤　実	14.	鈴木　隆
7.	斎藤　博	15.	高橋三郎
8.	渡辺　清		

女姓名で同姓同名の多いもの

1.	鈴木和子	9.	中村和子
2.	佐藤和子	10.	高橋幸子
3.	渡辺和子	11.	伊藤和子
4.	高橋和子	12.	佐藤幸子
5.	田中和子	13.	加藤和子
6.	鈴木幸子	14.	山田和子
7.	鈴木恵子	15.	斎藤和子
8.	小林和子		

5. この調査結果の利用について
銀行，証券，生保，損保等の業界では名寄せシステムを構築することが盛んに行われている。これにこの情報を有効に使うことを考えてみる。

(1) 1つの支店，支社で管理する顧客の規模は1万～5万件程度である。
この規模で同姓同名は約6％程度発生し1つの同姓同名は最大8名程度である。それ故漢字の名前と1桁～2桁の情報により個人を識別することが出来る。
1桁か2桁の情報には住所をコード化して使用すれば十分である。

(2) 全国 100万件程度の顧客を本社で一括管理していると仮定する。これを漢字の名前と何かの情報により名寄せを行うとする。この付加情報としては生年月日や全国の住所をコード化した住所コードが考えられる。1年は365日（366日）であるし，顧客の年令巾を20年とすると7,300種類にまで分類することが出来る。
7,300は181より十分大きいので漢字の姓名と生年月日によって個人の識別を行うことが出来る。
又別の方法として全国の住所をコード化した住所コードを利用することを考えてみる。全国の住所は丁目大字の段階までで約15万件ある。これは4～5桁の情報と同じである。181人とこれを比べてみるとかなり大きな差があります。
漢字の姓名と住所コードによれば十分個人を識別することができる。
（但しこの考えについては十分検討又はコンピュータによるシミュレーションを実施してみる必要がある）。

(3) 同姓同名がどのような名前で発生するかという資料があれば新人教育で教育資料として使うことができる。又支店，支社の事務指導においても有効に使うことができる。
これらの指導はほんのちよっとしたことかもしれないがこれにより大巾に事務の発生を防ぐことにもなる。

6. 仮名姓名の同姓同名の発生頻度
仮名姓名について同姓同名の調査を行った。調査内容を次に示す。

(i) 調査年月日：昭和52年4月
(ii) 調査データ：カナの姓名
(iii) 調査対象件数：926,145
(iv) データの性質：同一人のチェックが十分行われているファイル

(V) データの対象地区 ： 全 国

９２万件のデータをランダム・ナンバーを使用し１万件，３万件，４万件，９万件，１８万件，５０万件のデータに分割し分析を行った。

調査結果は次の表５，表６の通りである。

カナ姓名で同姓同名の発生頻度

	調査人数	同姓同名発生人数	%
1	11,380	664	5.83
2	33,307	4,478	13.43
3	46,297	9,002	19.44
4	92,699	25,387	27.38
5	185,415	68,924	37.17
6	463,081	240,304	51.89
7	926,145	584,986	63.16

表 5

カナ同姓同名最多発生人数

	調査人数	最多人数
1	11,380	5
2	33,307	9
3	46,297	14
4	92,699	30
5	185,415	57
6	463,081	143
7	926,145	266

表 6

カナの姓名と漢字の姓名を比べてみると同姓同名の発生割合が漢字にくらべ仮名は２倍ほど多いことがわかる。（１００万件の範囲内では）

このことからも漢字は仮名にくらべ個人の識別に充分役立つことがわかる。

カナ同姓同名と漢字同姓同名をグラフに表わすと図２のようになる。縦軸は％を表し横軸は対象データ件数を対数で表わし全体を半対数グラフで表わしたものである。

カナ，漢字データによる同姓同名の発生頻度

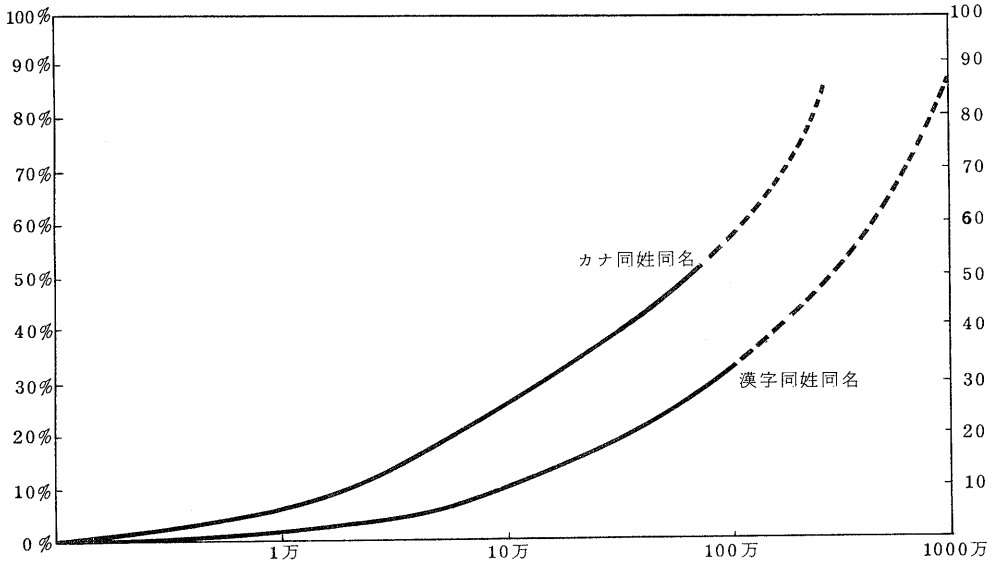


図 2

(-------- は予想
——— は実測)

１００万件以上は推定である。
このグラフからわかるように漢字姓名を用いても対象データが増加すれば個人の識別には限度があることが判る。
名寄せシステムをデザインする人々によって有効に利用されることを望みます。

7. おわりに

同姓同名というテーマはありふれたものであるが分析してみると大変興味深いものであった。この分析が色々な方面で有効に使われることを望みます。
この分析にあたって全国を均一にサンプリングしたデータが得られず一部かたよりがあると思います。これは今後の研究によって補正されることを期待します。
この分析のプログラム作成に協力してくれたシステム統括第１部滝沢章雄君，ＳＲＡ社 高山君に感謝致します。