Vol. 25 No. 4 情報処理学会論文誌 July 1984

収集語彙の母集団覆内率推定値の誤差の分散推定法の改善†

松岡 潤††*

仮名漢字変換システムや機械翻訳システムなどの自然言語処理システムにおいて，用語辞書などのテーブルウェアは，そのシステムの処理精度に大きく影響を与える一つの要因である．処理中の未格納語へのヒット率(テーブルウェアの覆外率と呼ぶ)が小さいことが強く要求される．無作為に集められた語の集合が母集団に対してもつ覆外率 $D_{0,N}$ の推定には，$D_{0,N}=C_{1,N+1}/(N+1)$ なる推定式[1]が用いられる．ここに $C_{1,N}$ は語彙調査で1回出現語の数であり，N は標本の大きさである．この推定式の誤差 w_N の分散 $V[w_N]$ については Robbins[4] の粗い評価式がある．また辞書の大きさに関係する異なり語数 k_N の分散 $V[k_N]$ については水谷の理論式[9]がある．本報告では語彙調査で得られる統計量だけから $V[w_N]$ および $V[k_N]$ を推定する方法を提案し，モンテカルロ法によって提案推定式の検証を行った．検証に用いた母集団の場合，提案推定式によって，覆外率の推定誤差の標準偏差 $\sigma[w_N]$ の精度は Robbins の評価式から1桁以上改善されること，また異なり語数の標準偏差 $\sigma[k_N]$ の上界は真値からの相対誤差が 20% 以下であることを示した．提案推定式の適用は ν 回出現語数 $C_{\nu,N}$ ($\nu\geq 2$) が $C_{1,N}$ を超えない程度の大きさの場合に限られるが，語の出現率分布関数の形にはよらない推定式である．

1. まえがき

計算機システムによる自然言語処理には，用語，用字等の辞書が不可欠である．ここにいう辞書は計算機内部に格納された辞書であり，自然言語処理プログラムが自由にアクセスできる形態のものである．たとえば仮名漢字変換システムにおける単語や文節単位の仮名文字列とそれの漢字仮名混じり表現との対応表などである．

そして自然言語処理システムの処理精度，処理性能は，辞書の良否に大きく依存する．

辞書を作成するためには，着目する自然言語の原始データを対象として，語彙調査が必要となる．語彙調査によって，辞書に収納すべき具体的な語と，それらの出現頻度，異なり語数等のデータが得られる．

しかし，このように語彙調査に基づいて作成した辞書を用いても，未登録の語に遭遇することがしばしば起こる．この遭遇の確率をその辞書の覆外率 (non-cover ratio) と名づける．覆外率は辞書の不完全性の一つの重要な指標と見られる．一方辞書の所要メモリの大きさは，各語に付随する属性等の項目が定まれば辞書の見出し語の数にほぼ比例する．辞書の完全性を一定水準以上にする上で必要となる見出し語の数は，語彙調査で現れた異なり語の数と直接かかわり合っている．

† An Advanced Evaluation Method for Tableware Cover Ratio Estimation Precision by HIROSHI MATSUOKA (Systems Development Laboratory, Hitachi, Ltd.).

†† (株)日立製作所システム開発研究所

* 現在 日立マイクロコンピュータエンジニアリング(株)

従来から収集語彙の母集団覆外率を求めるために，統計理論によって推定する方法[1),4),7),8),10)]と，直接測定を語彙収集手順の中に織り込んでゆく方法[9)]が研究されている．本報告では前者の立場に立ち，その推定誤差の分散を推定しようとする．従来の研究で，本報告の前提となる主要な事柄を以下に略記する．

無作為抽出で得られた延べ数 N の語の集合が母集団に対してもつ覆外率 $D_{0,N}$ は

$$E[D_{0,N}]=\frac{1}{N+1}E[C_{1,N+1}] \tag{1.1}$$

によって推定できることが Good[1)] によって示されている．ここに $E[x]$ は変量 x の期待値を表し，$C_{\nu,N}$ は無作為に抽出した大きさ N の標本において ν 回出現している語の数を表す．式(1.1)を覆外率簡易推定法と呼ぶこととする．式(1.1)を用いて $D_{0,N}$ を推定するときの誤差

$$w_N=\frac{C_{1,N+1}}{N+1}-D_{0,N} \tag{1.2}$$

の分散 $V[w_N]$ は

$$(N+1)V[w_N]=\sum_{i=1}^{L}g_i(1-g_i)^N\{1+(N-1)g_i\}-\sum_{i=1}^{L}\sum_{j\neq i}g_ig_j(1-g_i-g_j)^N \tag{1.3}$$

を満たすことが Robbins[4)]によって示された．ここに g_i は母集団における第 i 語の出現率を，L は母集団の語彙量を表す．また $V[x]$ は変量 x の分散を表す．Robbins は式(1.3)から，さらに

$$V[w_N] \leqq \frac{1}{N+1} \tag{1.4}$$

を導いた．本報告では式(1.4)より精度の高い$V[w_N]$の推定式(2.2)（後述）を提案する．

延べ語数 N の語彙調査で得られる異なり語数 k_N に関し，水谷[5]は

$$V[k_N]=\sum_{i=1}^{L}(1-g_i)^N-\left\{\sum_{i=1}^{L}(1-g_i)^N\right\}^2+\sum_{i=1}^{L}\sum_{j\neq i}(1-g_i-g_j)^N \tag{1.5}$$

と書けることを明らかにしている．

式(1.5)は母集団における各語の出現率 g_i を右辺に含んでいるが，g_i は一般にはわかっていない量であるので，この式から $V[k_N]$ を直接に算定することはできない．本報告では，語彙調査で得られる統計量から $V[k_N]$ を評価できる式(2.8)（後述）を提案する．

2章において $V[w_N]$ の推定式，および $V[k_N]$ の評価式の提案について説明を記し，3章において，精度の検証方法および結果について述べる．2章で示す式の証明は付録に示した．

2. 被覆率簡易推定法の誤差の分散 $V[w_N]$ および異なり語数の分散 $V[k_N]$ の推定法

2.1 $V[w_N]$ の推定法

本章において，次のことを前提条件とする．

（前提条件）「$\nu \geqq 2$ である $E[C_{\nu,N}]$ はすべて $E[C_{1,N}]$ と同等かそれより小さいものとする．」

式(1.2)による w_N について，

$$(N+1)^2V[w_N]=E[C_{1,N+1}]+2E[C_{2,N+2}]-\frac{1}{N+1}\{E[C_{1,N+1}]\}^2+o(N^{-1}\cdot E[C_{1,N}]) \tag{2.1}$$

が成り立つ．ただし，ここに $o(\varepsilon)$ は ε 程度以下の微小量を表す．本式の証明は付録(2)に示す．

この式の実用に当たっては，Nはかなり大きく（数十以上），かつ $E[C_{\nu,N+n}]-E[C_{\nu,N}]<n$ であるので，$E[C_{\nu,N+1}]$, $E[C_{\nu,N+2}]$ などを $C_{\nu,N}$ で置き換えて，

$$(N+1)^2V[w_N] \fallingdotseq C_{1,N}+2C_{2,N}-\frac{1}{N+1}(C_{1,N})^2 \tag{2.2}$$

としてよい．

2.2 $V[k_N]$ の近似算定法

まず，$V[k_N]$ について，既述前提条件の下に次の式が成り立つ．

$$V[k_N]=E[k_{2N}]-E[k_N]-\frac{1}{N}\{E[C_{1,N}]\}^2+\frac{1}{2N-1}E[C_{2,2N}]+o(N^{-2}\cdot\{E[C_{1,N}]\}^2) \tag{2.3}$$

この式の証明は付録(3)に示す．

この式の右辺は k_{2N}, $C_{2,2N}$ を含んでいるので，それらを大きさ N の標本の統計量で推定する必要がある．そのための式を以下に示す．

$$E[k_{2N}] \leqq E[k_N]+E[C_{1,N}]-\frac{2}{N+1}E[C_{2,N+1}] \tag{2.4}$$

が成り立つ．証明を付録(4)に示す．さらに

$$E[k_{2N}]>E[k_N]+E[C_{1,N}]\times\exp\left\{-\frac{2NE[C_{2,N+1}]}{(N+1)E[C_{1,N}]-2E[C_{2,N+1}]}\right\} \tag{2.5}$$

も成り立つことが証明される（付録(5)参照）．

$E[C_{2,2N}]$ に関しては，次式が成り立つ(付録(6))．

$$\frac{2(2N-1)}{N-1}E[C_{2,N}]\times\exp\left\{-\frac{3NE[C_{3,N+1}]}{(N+1)E[C_{2,N}]-3E[C_{3,N+1}]}\right\}<E[C_{2,2N}]<\frac{2N-1}{N-1}\frac{1}{e}E[C_{1,N}] \tag{2.6}$$

以上の式(2.3)～(2.6)によって，異なり語数の分散 $V[k_N]$ に関して，

$$E[C_{1,N}]\exp\left\{-\frac{2NE[C_{2,N+1}]}{(N+1)E[C_{1,N}]-2E[C_{2,N+1}]}\right\}-\frac{1}{N}\{E[C_{1,N}]\}^2+\frac{2E[C_{2,N}]}{N-1}\times\exp\left\{-\frac{3NE[C_{3,N+1}]}{(N+1)E[C_{2,N}]-3E[C_{3,N+1}]}\right\}<V[k_N]<\left\{1+\frac{1}{N-1}\frac{1}{e}\right\}E[C_{1,N}]-\frac{2}{N+1}E[C_{2,N+1}]-\frac{1}{N}\{E[C_{1,N}]\}^2 \tag{2.7}$$

が成り立つことがわかる．実用では，式(2.1)の直後に述べたと同様の理由により，

$$C_{1,N}\exp\left\{-\frac{2NC_{2,N}}{NC_{1,N}-2C_{2,N}}\right\}-\frac{(C_{1,N})^2}{N}+\frac{2}{N}C_{2,N}\exp\left\{-\frac{3NC_{3,N}}{NC_{2,N}-3C_{3,N}}\right\}<V[k_N]<\left(1+\frac{1}{Ne}\right)C_{1,N}-\frac{2}{N}C_{2,N}-\frac{(C_{1,N})^2}{N} \tag{2.8}$$

として適用することができる.

3. 精度の検証

3.1 検証の方法

前章で示した式(2.2)および式(2.8)の妥当性を検証する. 実際の語彙調査で, 同一の母集団から各種の大きさの標本を多数回抽出したデータはほとんどない. そこで語抽出シミュレーションによって検証することとした.

[標本の作成] まず, 各語の出現率 g_i ($i=1, 2, \cdots, L$) のわかっている十分大きな語の集合をとり, これを実験上の母集団とする. この母集団から任意の大きさの標本を抽出するのに, 次の方法を用いる.

$$\begin{cases} f(j)=\sum_{i=1}^{j} g_i \quad (j=1, 2, \cdots, L) & (3.1a) \\ f(0)=0 & (3.1b) \end{cases}$$

なる関数を考えると $f(L)=1$ である. 区間 $[0, 1)$ の値域をもつ一様乱数を発生させる. いまその一つを R とする.

$$f(j_R-1) \leqq R < f(j_R) \qquad (3.2)$$

を満足する整数 j_R を求める. この j_R を, 抽出された一つの語の番号と見なす. 乱数発生を繰り返し, N 個の抽出語の集合を作成する. これが標本である.

[標本の覆外率等の計算] 作成される大きさ N の標本について, 次のものを求める.

(a) 異なり語数 k_N

(b) 覆外率 $D_{0,N}$

(c) 覆外率簡易推定法の誤差 $w_N^* = \dfrac{C_{1,N}}{N+1} - D_{0,N}$

(d) ν 回出現語数 $C_{\nu,N}$

標本の記録には, 語の番号 j_R と, その語の出現した回数 r_R とを記録してゆくようにする. 標本作成過程で新たに一つの j_R を得た段階では, まずすでに抽出されている語の番号のなかに j_R と等しい値のものがあるかどうかを調べ, あればその r_R に1を加え, なかったときに新たに j_R を記録し, かつその r_R を1とする. これにより(a)異なり語数 k_N は記録された j_R の個数である. (d) ν 回出現語数 $C_{\nu,N}$ は $r_R=\nu$ なる記録の総数である. また覆内率 s_N は, 記録されたすべての j_R について g_{j_R} の和をとる. すなわち

$$s_N = \sum g_{j_R} \quad \begin{pmatrix} r_R \geqq 1 \text{ なるすべての} \\ \text{記録についての和} \end{pmatrix} \qquad (3.3)$$

によって算出できる. (b)覆外率 $D_{0,N}$ はこの s_N によって

$$D_{0,N}=1-s_N \qquad (3.4)$$

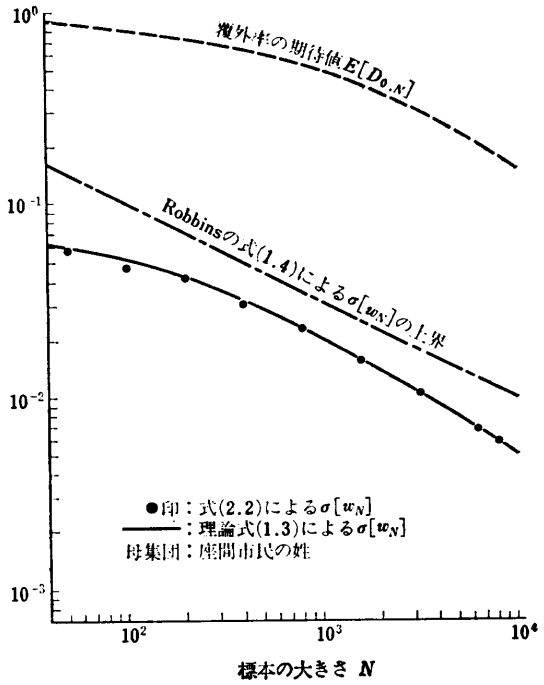

図 1 覆外率簡易推定法の誤差 w_N の分散推定式(2.2)の検証
Fig. 1 Tests of formula (2.2), the error variance of formula (1.1).

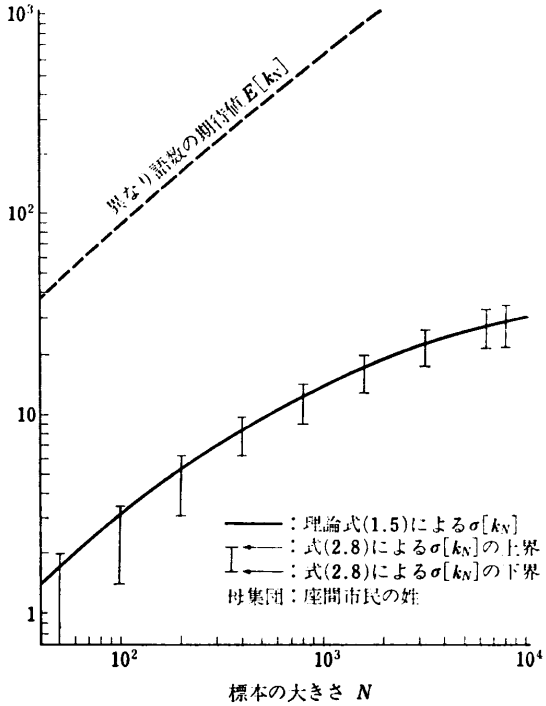

図 2 異なり語数 k_N の分散推定式(2.8)の検証
Fig. 2 Tests of formula (2.8), the variance of number of different words k_N.

表 1 母集団

Table 1 The population used for the tests.

(a) 出現回数 ν	(b) ν 回出現語の数 C_ν	(c) ν 回出現語一つの出現率 g	(a) 出現回数 ν	(b) ν 回出現語の数 C_ν	(c) ν 回出現語一つの出現率 g
449	1	$1.9397(\times10^{-2})$	41	1	$1.7712(\times10^{-3})$
436	1	1.8835	40	3	1.7280
343	1	1.4818	39	4	1.6848
291	1	1.2571	38	3	1.6416
253	1	1.0930	37	1	1.5984
228	1	$9.8497(\times10^{-3})$	36	5	1.5552
217	1	9.3745	35	3	1.5120
207	1	8.9425	34	5	1.4688
200	1	8.6401	33	3	1.4256
178	1	7.6896	32	1	1.3824
172	1	7.4304	31	6	1.3392
148	1	6.3936	30	5	1.2960
147	1	6.3504	29	5	1.2528
136	1	5.8752	28	1	1.2096
131	1	5.6592	27	9	1.1664
129	2	5.5728	26	8	1.1232
112	1	4.8384	25	7	1.0800
102	1	4.4064	24	6	1.0368
98	1	4.2336	23	3	$9.9361(\times10^{-4})$
91	1	3.9312	22	13	9.5041
86	2	3.7152	21	8	9.0721
84	2	3.6288	20	15	8.6401
82	1	3.5424	19	10	8.2081
81	1	3.4992	18	8	7.7760
79	4	3.4128	17	14	7.3440
78	1	3.3696	16	21	6.9120
74	2	3.1968	15	17	6.4800
72	1	3.1104	14	14	6.0480
71	1	3.0672	13	31	5.6160
70	1	3.0240	12	25	5.1840
67	1	2.8944	11	24	4.7520
66	1	2.8512	10	48	4.3200
65	1	2.8080	9	54	3.8880
63	2	2.7216	8	50	3.4560
62	1	2.6784	7	81	3.0240
61	1	2.6352	6	112	2.5920
60	1	2.5920	5	117	2.1600
57	1	2.4624	4	218	1.7280
56	2	2.4192	3	315	1.2960
55	1	2.3760	2	754	$8.6401(\times10^{-5})$
54	1	2.3328	1	3569	4.3200
53	3	2.2896			
51	3	2.2032			
50	4	2.1600			
49	3	2.1168			
48	2	2.0736			
47	5	2.0304			
46	3	1.9872			
43	4	1.8576			
42	1	1.8144			

母集団：座間市電話帳による座間市民の姓
総語数：23,148, 異なり語数：5,675

表 2 抽出された標本の例
Table 2 Examples of sample.

項目		標本の大きさ N								
		50	100	200	400	800	1,600	3,200	6,400	8,000
(a) 異なり語数	k_N	48	93	169	300	523	883	1,440	2,264	2,551
(b) 覆外率	$D_{0,N}$	0.88135	0.78525	0.72870	0.62914	0.52702	0.41562	0.30730	0.20430	0.17791
(c) 簡易推定法の誤差	$w_N^*\times 100$	3.8648	8.4747	1.1298	−0.1635	−2.3274	−1.3750	−0.7298	0.8201	0.1838
(d)	$C_{1,N}$	46	87	148	251	403	643	960	1,360	1,438
ν 回出現語数 $C_{\nu,N}$	$C_{2,N}$	2	5	16	30	71	132	226	396	484
($\nu\leqq 5$)	$C_{3,N}$	0	1	2	9	22	41	82	157	189
	$C_{4,N}$	0	0	2	4	5	21	62	91	102
	$C_{5,N}$	0	0	0	2	9	11	28	60	69

注) $C_{\nu,N}$ ($\nu\geqq 6$) については記載を省略した.

表 3 式(2.2)による $\sigma[w_N]$ 推定値の誤差
Table 3 Erros of $\sigma[w_N]$ by formula (2.2).

項目		標本の大きさ N								
		50	100	200	400	800	1,600	3,200	6,400	8,000
(1) 1回出現語数	$C_{1,N}$	46	87	148	251	403	643	960	1,360	1,438
(2) 2回出現語数	$C_{2,N}$	2	5	16	30	71	132	226	396	484
(3) 式(2.2)による推定値	$(N+1)^2V[w_N]$	8.510	22.059	71.025	153.890	342.242	648.756	1,124.090	1,863.045	2,147.552
(4) 上記(3)による $\sqrt{V[w_N]}$	δ	0.0572	0.0465	0.0419	0.0309	0.0231	0.0159	0.0105	0.0067	0.0058
(5) 理論式(1.3)による $\sqrt{V[w_N]}$	σ	0.0591	0.0508	0.0408	0.0312	0.0227	0.0158	0.0105	0.0068	0.0058
(6) δ の相対誤差 $(\delta-\sigma)/\sigma$	ε_σ (%)	−3.2	−8.5	2.7	−1.0	1.8	0.6	0.0	−1.5	0.0
(7) Robbins による $\sqrt{V[w_N]}$ 上限	σ_R	0.1400	0.0995	0.0705	0.0499	0.0353	0.0250	0.0177	0.0125	0.0112
(8) σ_R の相対誤差 $(\sigma_R-\sigma)/\sigma$	ε_R (%)	136.9	95.9	72.8	59.9	55.5	58.2	68.6	83.8	93.1

である. (c)覆外率簡易推定法の誤差 w_N は, 本来, 式(1.2)であるが $C_{1,N+1}$ を $C_{1,N}$ で置き換えた実用の形 (これを w_N^* と記すこととする) で求めた. これは(b)と(d)の $C_{1,N}$ から簡単に計算される.

[理論値の計算] 上記の語抽出のシミュレーションとは別に, 式(1.3)および式(1.5)によって $V[w_N]$ および $V[k_N]$ の理論値を計算した.

[実験用の母集団] 実験用の母集団としては座間市民の姓の集合を採用した. そのデータは電話帳[6]に記載されている座間市民の姓を筆者が数えたものであり, **表1**のとおりであった. 同表の(c)欄の値は(a)欄の値の総語数に対する比として算出したものである.

3.2 検 証

(1) 作られた標本例

各種サイズの標本の一例を示すと**表2**のとおりである.

(2) 覆外率簡易推定法の誤差の分散 $V[w_N]$ の検証

表2の標本の $C_{\nu,N}$ を用いて, 式(2.2)による $V[w_N]$ を算出したものと, 式(1.3)による理論値との比較を**図1**に示す. ただし見やすさのために分散の2乗根である標準偏差 $\sigma[w_N]$ の形で表している. 同図には Robbins の式(1.4)による計算値をも示した. 式(2.2)による推定値が Robbins の式よりも格段に精度が高いことがわかる.

(3) 異なり語数の分散 $V[k_N]$ の検証

表2の標本の $C_{\nu,N}$ を用いて, 式(2.8)によって $V[k_N]$ の上界・下界を算出する. これと, 理論式(1.5)との比較を**図2**に示す. $V[w_N]$ の場合に比較して精度は落ちるが, $V[k_N]$ の大きさの程度を式(2.8)がよく表していることがわかる.

表 4　式(2.8)による $\sigma[k_N]$ の誤差
Table 4　Errors of $\sigma[k_N]$ by formula (2.8).

項目		標本の大きさ N								
		50	100	200	400	800	1,600	3,200	6,400	8,000
(1) 1回出現語数	$C_{1,N}$	46	87	148	251	403	643	960	1,360	1,438
(2) 2回出現語数	$C_{2,N}$	2	5	16	30	71	132	226	396	484
(3) 3回出現語数	$C_{3,N}$	0	1	2	9	22	41	82	157	189
(4) 式(2.8)による $\sqrt{V[k_N]}$ の上界	σ_s	2.0	3.4	6.2	9.7	14.1	19.6	25.9	32.7	34.3
(5) 理論式(1.5)による $\sqrt{V[k_N]}$	σ_k	1.7	3.1	5.3	8.3	12.2	17.1	22.4	27.5	28.7
(6) σ_s の相対誤差 $(\sigma_s-\sigma_k)/\sigma_k$	ε_s(%)	17.6	9.7	17.0	16.9	15.6	14.6	15.6	18.9	19.5
(7) 式(2.8)による $\sqrt{V[k_N]}$ の下界	σ_I	0.0	1.4	3.1	6.3	9.0	13.0	17.7	21.7	21.8
(8) σ_I の相対誤差 $(\sigma_I-\sigma_k)/\sigma_k$	ε_I(%)	−100.0	−54.8	−41.5	−24.1	−26.2	−24.0	−21.0	−21.1	−24.0

表 5　標本の大きさ N が 1,600 の標本の例および式(2.2)の適用
Table 5　Examples of sample (size N=1,600) and applications of formula (2.2).

	項目		標本の例 No. 1	2	3	4
標本の統計量	(a) 異なり語数	k_N	900	884	919	877
	(b) 覆外率	$D_{0,N}$	0.40955	0.40646	0.40767	0.42167
	(c) 簡易推定法の誤差	w_N^*	0.008578	−0.006458	0.021706	−0.022922
	(d-1)					
		$C_{1,N}$	669	640	687	638
		$C_{2,N}$	121	127	118	115
		$C_{3,N}$	45	48	49	50
	ν 回出現語数	$C_{4,N}$	21	21	17	23
	$C_{\nu,N}$	$C_{5,N}$	6	14	14	16
	($\nu \leqq 10$)	$C_{6,N}$	8	7	7	5
		$C_{7,N}$	7	4	4	7
		$C_{8,N}$	3	3	4	3
		$C_{9,N}$	4	4	1	3
		$C_{10,N}$	3	4	2	3
	(d-2)		ν $C_{\nu,N}$	ν $C_{\nu,N}$	ν $C_{\nu,N}$	ν $C_{\nu,N}$
	$\nu>10$ の $C_{\nu,N}$		12 4	11 4	11 1	12 4
	(記載のない ν については $C_{\nu,N}=0$)		13 1	17 1	12 2	15 4
			17 1	20 2	13 3	18 1
			18 1	22 1	14 1	21 1
			19 2	23 1	15 1	22 2
			21 1	26 1	16 1	27 1
			28 1	27 1	17 4	32 1
			31 1	39 1	18 1	
			39 1		22 1	
					34 1	
(e) 式(2.2)による $\sqrt{V[w_N]}$		δ	0.01570	0.01578	0.01566	0.01547
(f) δ の相対誤差 $(\delta-\sigma)/\sigma$		ε_δ(%)	−0.7	−0.1	−0.9	−2.1
(g) 式(2.8)による $\sqrt{V[k_N]}$ 上界		σ_s	19.73	19.60	19.80	19.59
(h) σ_s の相対誤差 $(\sigma_s-\sigma_k)/\sigma_k$		ε_s(%)	15.4	14.6	15.8	14.6

母集団：座間市民の姓

3.3 結果の検討

近似式(2.2)および(2.8)の精度がどの程度であるかを調べる．表2に示した標本について，式(2.2)の推定値が，理論計算値に対してもつ相対誤差は，**表3**の(6)欄に示すとおりである．誤差の大きいところでも10% 以下である．Robbins の評価式を用いた場合の相対誤差をも同表(8)欄に示した．本稿の推定式(2.2)で精度が1桁以上改善されたことがわかる．

式(2.8)による推定値の相対誤差は，**表4**に示した．実用上大切な上界の相対誤差は 10〜20% 程度であることがわかる．精度は高いとはいえないが，語彙調査で得られるデータから簡単に $V[k_N]$ を求める方法が得られたことは有意義と考える．

標本の違いに対する推定値の安定性を確認するために，同一の大きさの標本を数例作ってみた．その結果を**表5**に示す．式(2.2)による w_N の標準偏差 $\bar{\sigma}$ と，その理論値(すでに表3(5)欄に示してある)に対する相対誤差 ε_σ とを同表(e)，(f)欄に示す．標本が異なっても $\bar{\sigma}$ の相対誤差が数パーセント以内であることがわかる．また式(2.8)による k_N の標準偏差の上界 σ_s についても同表(g)，(h)欄に示す．標本の違いの σ_s，ε_s への影響が小さいことが示されている．

4. むすび

語彙調査で無作為抽出して得られる標本の覆外率を簡易推定法で推定したときの誤差の分散 $V[w_N]$，および標本の異なり語数の分散 $V[k_N]$ を，語彙調査で得られる統計量から推定する方法を提案した．また語抽出シミュレーションで得られたデータによって推定精度の検証を行い，$\sigma[w_N]$ については悪いところでも相対誤差 10% 以内であることを，また $\sigma[k_N]$ について推定上界の真値との相対誤差は 20% 程度以下であることを示した．提案した方法は，母集団の語彙の出現率分布関数の形によらない方法である．

謝辞 覆外率簡易推定式に関し，文献4)を紹介いただいた慶応大学渋谷政昭教授に，また本研究の機会を与えていただいた当社コンピュータ事業本部三浦武雄本部長，同システム開発研究所川崎淳所長に感謝の意を表します．

参考文献

1) Good, I. J.: The Population Frequencies of Species and the Estimation of Population Parameters, *Biometrika*, Vol. 40, Parts 3 & 4, pp. 237-264 (1953).
2) 森口繁一他：数学公式II，p. 340，岩波書店，東京 (1957).
3) フェラー，W.（河田龍夫監訳）：確率論とその応用，p. 274，紀伊国屋書店，東京 (1963).
4) Robbins, H. E.: Estimating the Total Probability of the Unobserved Outcomes of an Experiment, *Ann. Math. Stat.*, Vol. 39, No. 1, pp. 256-257 (1968).
5) 水谷静夫：国語学五つの発見再発見，p. 154，創文社，東京 (1974).
6) 日本電信電話公社：神奈川県北部版 50音別電話帳 昭和51年11月1日現在，p. 558，日本電信電話公社 (1977).
7) 松岡，新田：語彙出現率に関する推定法，日本語情報処理シンポジウム，pp. 325-331 (1978).
8) 渋谷政昭：多用漢字の選出，計量国語学，Vol. 11, No. 7, pp. 322-323 (1978).
9) 田中康仁，長田孝治，土屋信一：科学技術文献抄録における片仮名列の解析，計量国語学，Vol. 14, No. 1, pp. 15-20 (1983).
10) 木村睦子：辞書収録語彙とカバー率，計量計画研究所研究報告 '82, pp. 45-52 (1983).

付録

(1) 用語・記号の説明および前提

覆内率：語の母集団と，語の集合 α とがあるとする．母集団から無作為に語をとり出したとき，その語がすでに α に含まれている確率の期待値を，その母集団に対する α の覆内率という．

覆外率：=1−覆内率

$D_{\nu,N}$：母集団から無作為に抽出した大きさ N の標本において，ν 回出現している語の集合の母集団に対する覆内率．$\nu \geqq 0$, $N \geqq 1$.

以上の定義の下に，次の式が成り立つ．

$$\sum_{i=1}^{L} g_i = 1 \tag{A1.1}$$

$$E[k_N] = \sum_{\nu=1}^{N} E[C_{\nu,N}] = \sum_{i=1}^{L} \{1-(1-g_i)^N\} \tag{A1.2}$$

$$\sum_{\nu=0}^{N} E[D_{\nu,N}] = 1 \tag{A1.3}$$

$$E[C_{\nu,N}] = \sum_{i=1}^{L} \binom{N}{\nu} g_i{}^{\nu}(1-g_i)^{N-\nu} \tag{A1.4}$$

$$E[D_{\nu,N}] = \sum_{i=1}^{L} \binom{N}{\nu} g_i{}^{\nu+1}(1-g_i)^{N-\nu} \tag{A1.5}$$

$$E[D_{\nu,N}] = \frac{\nu+1}{N+1} E[C_{\nu+1,N+1}] \tag{A1.6}$$

$$(\nu+1)E[C_{\nu+1,N+1}] + (N-\nu+1)E[C_{\nu,N+1}] - (N+1)E[C_{\nu,N}] = 0 \tag{A1.7}$$

式(A1.1)～式(A1.3)は明らかであり，また式(A1.4)～式(A1.7)についてはすでに報告している[7]ので説明を省略する．

(2)　式(2.1)の証明

(i)　まず次の（予備定理1）から証明する．

《予備定理1》　$1\leqq\nu\leqq N$ において，次の式が成り立つ．

$$E[C_{2\nu,2N}]\leqq\frac{\binom{2N}{2\nu}}{\binom{N}{\nu}}\left(\frac{\nu}{N}\right)^{\nu}\left(1-\frac{\nu}{N}\right)^{N-\nu}E[C_{\nu,N}] \tag{A2.1}$$

（証明）　x の関数 $x^{\nu}(1-x)^{N-\nu}$ は値域 $[0,1]$ では $x=\nu/N$ において最大値をとることから明らかである．　（予備定理1　証終り）

右辺の係数 T を評価する．$\nu=N$ においては $T=1$．$\nu<N$ については，自然数 n に関して

$$(2\pi)^{1/2}n^{n+1/2}e^{-n+1/(12n+1)}<n!<(2\pi)^{1/2}n^{n+1/2}e^{-n+1/12n}$$

である[3]ことを援用することにより

$$T<2^{-1/2}\exp\left\{\frac{1}{24\nu}-\frac{1}{24N}+\frac{1}{24(N-\nu)}+\frac{1}{144N^2}+\frac{1}{576(N-\nu)^2}+\frac{1}{576\nu^2}\right\}<0.78 \tag{A2.2}$$

である．とくに $1=\nu<N$ の場合は

$$T<\frac{2N-1}{N-1}\cdot\frac{1}{e} \tag{A2.3}$$

(ii)　本題に戻る．式(1.3)右辺第2項に着目する．

$$I_N=\sum_{i=1}^{L}\sum_{j\neq i}g_ig_j(1-g_i-g_j)^N \tag{A2.4a}$$

$$J_N=\sum_{i=1}^{L}\sum_{j=1}^{L}g_ig_j(1-g_i-g_j)^N \tag{A2.4b}$$

$$L_N=\sum_{i=1}^{L}g_i^2(1-2g_i)^N \tag{A2.4c}$$

とおくと $I_N=J_N-L_N$ である．

$$(1-g_i-g_j)^N=\sum_{r=0}^{N}\binom{N}{r}(-)^r(1-g_i)^{N-r}(1-g_j)^{N-r}g_i^rg_j^r \tag{A2.5}$$

であるから，式(A1.5)および式(A1.6)を援用することにより，

$$J_N=\sum_{r=0}^{N}\frac{(-)^r}{\binom{N+1}{r+1}}\frac{r+1}{N+1}\{E[C_{r+1,N+1}]\}^2 \tag{A2.6}$$

が得られる．次に L_N については，

$$(1-2g_i)^N=\sum_{r=0}^{N}\binom{N}{r}(-)^r(1-g_i)^{2(N-r)}g_i^{2r} \tag{A2.7}$$

であるので，J_N の場合と類似の置きかえによって，

$$L_N=\sum_{r=0}^{N}\frac{(-)^r\binom{N}{r}}{\binom{2N+1}{2r+1}}\frac{r+1}{N+1}E[C_{2r+2,2N+2}] \tag{A2.8}$$

と書ける．式(1.3)右辺第1項についても式(A1.5)，式(A1.6)による置きかえを施すことにより，同式は

$$(N+1)^2V[w_N]=E[C_{1,N+1}]+\frac{2(N-1)}{N+2}E[C_{2,N+2}]-(N+1)J_N+(N+1)L_N \tag{A2.9}$$

と書かれる．右辺第2項は，2.1節の前提条件を考慮することにより，$2E[C_{2,N+2}]+o(N^{-1}E[C_{1,N}])$ と書かれる．第3項は式(A2.6)から $\{E[C_{1,N+1}]\}^2/(N+1)+o(N^{-2}\{E[C_{1,N+1}]\}^2)$ であるが，ここの微小量は第2項のものに含めて表現できる．なお一般に g_i は最大でも数パーセントという小さい数であるので，式(A1.4)から $E[C_{N,N}]$，$E[C_{N-1,N}]$ はほとんど0である．第4項は式(A2.8)から $o(N^{-1}E[C_{2,2N}])$ であるが（予備定理1）により $o(N^{-1}E[C_{1,N}])$ としてよい．よって式(A2.9)は式(2.1)と一致する．（証明終り）

(3)　式(2.3)の証明

式(1.5)は次のように変形できる．

$$V[k_N]=\sum_{i=1}^{L}\{1-(1-g_i)^N\}-\left[\sum_{i=1}^{L}\{1-(1-g_i)^N\}\right]^2+\sum_{i=1}^{L}\sum_{j\neq i}\{1-(1-g_i)^N-(1-g_j)^N+(1-g_i-g_j)^N\} \tag{A3.1}$$

右辺第3項を M とおき，式(A1.2)を援用することにより，この式は

$$V[k_N]=E[k_N]-(E[k_N])^2+M \tag{A3.2}$$

と書ける．M は，

$$M=\sum_{i=1}^{L}\sum_{j=1}^{L}\{1-(1-g_i)^N-(1-g_j)^N+(1-g_i-g_j)^N\}-\sum_{i=1}^{L}\{1-2(1-g_i)^N+(1-2g_i)^N\} \tag{A3.3}$$

と変形できる．式(A2.5)により，式(A3.3)の右辺第1項は

$$\sum_{i=1}^{L}\sum_{j=1}^{L}\{1-(1-g_i)^N-(1-g_j)^N+(1-g_i)^N(1-g_j)^N+\sum_{r=1}^{N}\binom{N}{r}(-)^r(1-g_i)^{N-r}(1-g_j)^{N-r}g_i^r g_j^r\}=\Big[\sum_{i=1}^{L}\{1-(1-g_i)^N\}\Big]^2+\sum_{r=1}^{N}\frac{(-)^r}{\binom{N}{r}}\{E[C_{r,N}]\}^2$$

と変形できる．上の右辺第1項は $\{E[k_N]\}^2$ と書ける．また式(A2.7)により，式(A3.3)の右辺第2項は，負符号を除き

$$\sum_{i=1}^{L}\Big\{1-2(1-g_i)^N+(1-g_i)^{2N}+\sum_{r=1}^{N}\binom{N}{r}(-)^r(1-g_i)^{2(N-r)}g_i^{2r}\Big\}=2E[k_N]-E[k_{2N}]+\sum_{r=1}^{N}\frac{\binom{N}{r}(-)^r}{\binom{2N}{2r}}E[C_{2r,2N}]$$

と書けるので，これらの関係を式(A3.2)に持ち込むことにより，

$$V[k_N]=E[k_{2N}]-E[k_N]+\sum_{r=1}^{N}\frac{(-)^r}{\binom{N}{r}}\{E[C_{r,N}]\}^2-\sum_{r=1}^{N}\frac{(-)^r\binom{N}{r}}{\binom{2N}{2r}}E[C_{2r,2N}]\qquad(A3.4)$$

が得られる．前提条件を考慮すれば，右辺第3項の和は $\{E[C_{1,N}]\}^2/N+o(N^{-2}\{E[C_{1,N}]\}^2)$ と書け，右辺第4項は（予備定理1）をも援用することにより，$(2N-1)^{-1}E[C_{2,2N}]+o(N^{-2}E[C_{1,N}])$ と書ける．後者の微小量は前者のそれに含まれるので式(2.3)が得られる．（証明終り）

（4） 式(2.4)の証明

まず，次の（予備定理2）から始める．

《予備定理2》 $N\geqq 2$ において次式が成り立つ．

$$E[k_N]=\sum_{r=1}^{N-1}E[D_{0,r}]+1\qquad(A4.1)$$

（予備定理2の証明）式(A1.2)より $E[k_N]-E[k_{N-1}]=\sum_{i=1}^{L}\{1-(1-g_i)^N\}-\sum_{i=1}^{L}\{1-(1-g_i)^{N-1}\}$ が得られ，式(A1.5)からこれは $E[D_{0,N-1}]$ に等しい．これと，$E[k_1]=1$ であることにより式(A4.1)が成り立つ．（予備定理2 証終り）

《予備定理3》 次の式が成り立つ．

$$E[k_{2N}]\leqq E[k_N]+\frac{N}{N+1}E[C_{1,N+1}]\qquad(A4.2)$$

（予備定理3の証明）（予備定理2）から $E[k_{2N}]-E[k_N]=\sum_{n=N}^{2N-1}E[D_{0,n}]$ である．明らかに，$E[D_{0,N}]\geqq E[D_{0,N+1}]\geqq\cdots\geqq E[D_{0,2N-1}]$ であるからこの和は $N\cdot E[D_{0,N}]=N\cdot E[C_{1,N+1}]/(N+1)$ を越えない．（予備定理3 証終り）

本題に戻る．式(A4.2)の右辺第2項に対し，式(A1.7)で $\nu=1$ の場合を代入することにより，式(2.4)が得られる．（証明終り）

（5） 式(2.5)の証明

《予備定理4》 次の式が成り立つ．

$$\frac{E[C_{\nu,2N}]}{E[C_{\nu,N}]}>\frac{\binom{2N}{\nu}}{\binom{N}{\nu}}\times\exp\left\{-\frac{N(\nu+1)E[C_{\nu+1,N+1}]}{(N+1)E[C_{\nu,N}]-(\nu+1)E[C_{\nu+1,N+1}]}\right\}\qquad(A5.1)$$

（予備定理4の証明）式(A5.1)左辺を U とおき，また

$$a_i=\frac{g_i^\nu(1-g_i)^{N-\nu}}{\sum_{j=1}^{L}g_j^\nu(1-g_j)^{N-\nu}}$$

とおけば，

$$U=\frac{\binom{2N}{\nu}}{\binom{N}{\nu}}\sum_{i=1}^{L}a_i(1-g_i)^N>\frac{\binom{2N}{\nu}}{\binom{N}{\nu}}\sum_{i=1}^{L}a_i\exp\left\{-\frac{Ng_i}{1-g_i}\right\}\qquad(A5.2)$$

と書かれる．$\sum_{i=1}^{L}a_i=1$ と $\exp\{-Nx/(1-x)\}$ が下に凸な関数であることから

$$U>\frac{\binom{2N}{\nu}}{\binom{N}{\nu}}\exp\left\{-\frac{N\sum_{i=1}^{L}a_ig_i}{1-\sum_{i=1}^{L}a_ig_i}\right\}\qquad(A5.3)$$

が成り立つ（Jensen の定理[2]）．（予備定理4 証終り）

本題に戻る．（予備定理2）から $E[k_{2N}]-E[k_N]=\sum_{n=N}^{2N-1}E[D_{0,n}]$ であるが，（予備定理3）の証明で述べたことを考慮することにより，これは $N\cdot E[D_{0,2N-1}]$

を下まわらない．さらに式(A1.6)により $E[k_{2N}]-E[k_N]\geqq E[C_{1,2N}]/2$ と書かれる．この右辺に（予備定理4）の関係を持ち込むことにより，式(2.5)の成り立つことがわかる．（証明終り）

（6） 式(2.6)の証明

式(2.6)の右側の不等号は（予備定理1）により明らかであり，左側の不等号は（予備定理4）により明らかである．（証明終り）

(昭和58年6月 8 日受付)
(昭和59年2月14日採録)