研究ノート

著者推定問題における統計的手法

統計数理研究所　村　上　征　勝

(1978年12月　受付)

1. はじめに

洋の東西を問わず，著名な思想家，作家の著書として愛読され続けてきた作品の中にも，その成立について疑問がもたれているものが少なからず存在する．勿論，これらの疑問を解決すべく，それぞれの分野固有の伝統的な方法によって，内容，成立等に関して多くの研究がなされてきた．しかし，これとは別に作品の数量的な性質に注目して，この種の問題の解決を試みた研究者達もいる．この稿では，彼等が用いた研究方法やその成果を統計学の観点から要約，整理し，内在する問題点を明確にする．第2節では著者推定問題と密接に関連する執筆順序推定問題を，第3節では代表的な著者推定問題を紹介し，最後の第4節で統計的パターン認識の立場から問題点を考察する．

2. 執筆順序推定問題

西暦500年～1000年頃，パレスチナの地において旧約聖書の研究に従事していたマソラ学者(Masoretes) 達は，ヘブライ語の旧約聖書の索引作りの仕事の一部として，聖書各巻の文字数，語数，宗教的な意味を持つ語の頻度などを調べていた．文献の数量的な性質に対するこのような興味は19世紀に入り，執筆順序の推定や著者の推定に関する研究へと発展する．そして，前者における主な対象はプラトンとシェクスピアであった．

1867年，スコットランドの L. Campbell はプラトンの著作の執筆順序の推定に関する研究を発表した [1]．彼はプラトンの絶筆とされている「法律」を手がかりとして，主に哲学的用語について晩年の6つの対話篇の語彙を「法律」の語彙と比較して，その著作年代を推定した．この研究を契機に，その後，W. Dittenberger (1896年) [2], W. Lutosławski (1897年) [3], C. Ritter (1935年) [4], Cox と Brandwood (1959年) [5] 等によってプラトンの執筆順序の推定が，また C. Bathurst (1857年) [6], J. K. Ingram (1874年) [7], F. G. Fleay (1874年) [8], F. J. Furnival (1887年, 1907年) [9], [10], G. König (1888年) [11], H. Conrad (1897年) [12], E. K. Chambers (1930年) [13], M. R. Yardi (1946年) [14] 等によってシェクスピアの著作の執筆順序の推定がおこなわれた．シェクスピアについては，生年月日を始めとして，彼の生涯や作品について謎の部分が多く，また当時，合作や劇団所有の古い脚本や他人の作品の修正加筆が盛んであったこともあって，著者推定問題としても興味あるものが少なくない* が，さらに一部の人々はシェクスピアの存在までも疑いを抱いていた（第3節のA参照）．ところで，これらの研究の多くは語彙の比較がその中心であり必ずしも統計が重要な役割をはたしてはいない．統計的な意味で，また著者推定問題との関連で興味を引くのは Yardi の研究である．彼はシェクスピアの各作品中の "full split line", "redundant final

---

* 疑問がもたれている作品としては「ヘンリー六世」,「タイタス・アンドロニカス」, 合作の疑いのある作品としては「じゃじゃ馬ならし」,「トロイダスとクレラダ」,「アテネのタイモン」,「ペリクリーズ」,「ヘンリー八世」があげられている

syllable" "unsplit line with pause" の割合を基に，各作品について彼のいうところの識別関数値 u を計算し，u とシェクスピアの処女作品が執筆された年から，ある作品が書かれた年までの経過年数 v との間には近似的に

$$u = 0.7204 + 0.07645v$$

なる関係があると報告している．

3. 著者推定問題

A. モードによる推定

著者推定問題における統計学の有用性を最初に示唆したのは Augustes de Morgan である．1851年，彼はケンブリッジ大学時代の友人にあてた手紙の中で，単語の長さの平均値に作家の特徴が表われるであろうと記している [15] [16]．この de Morgan の考えにヒントを得たオハイオ州立大学の地球物理学者 T. C. Mendenhall は，1887年に Science 誌に発表した論文 [17] の中で Thackeray, Dickens, Mill 等の文章に現われる単語の長さの度数分布——彼はこの度数分布を word spectrum と呼んだ——を示し，著者推定問題に対する word spectrum の有効性を説いた．1901年，彼はシェクスピアとフランシス・ベーコンの word spectrum を実に40万語，20万語という膨大なデータから求め [18]，シェクスピアのモードは4文字から成る単語であり，ベーコンのモードは3文字から成る単語であることを示すことによって (図1)，当時一部の人々によって信じられていた，シェクスピアなる人物は歴史上に存在せず，ベーコンが圧政抗議のため一連の風刺劇を書いたという説を否定した．これが，著者推定問題に統計が主役となって登場した最初の研究である．

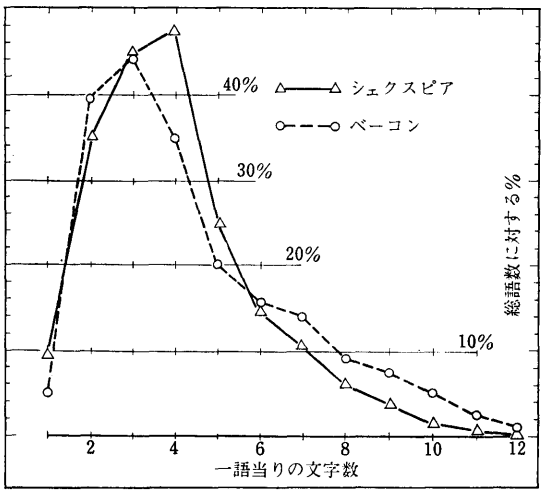


図1 シェクスピア，ベーコンの語の長さの分布

B. 平均値，中央値等による推定

1939年，統計学者 Undy Yule は，Bacon, Coleridge, Lamb, Macauley の文の長さを調査し，その平均値，中央値，四分位範囲等は作家間で異なるが，同一作家の作品では安定した値となるという結果を得た．彼はこの結果をもとに "The Imitation of Christ (原題名 De Imitatione Christi キリストにならいて)" と "Natural and Political Observations···made upon the Bills of Mortality (死亡表に関する…… 自然的および政治的諸観察)" の著者推定をこれらの

表 1. The Imitation of Christ と à Kempis, Gerson の比較

	The Imitation of Christ	à Kempis	Gerson 1	Gerson 2
平均値	16.2 語	17.9 語	23.4 語	22.7 語
中央値	13.8	15.1	19.6	18.9
四分位範囲	9.2	11.8	18.8	17.8

表 2. Graunt と Petty の比較

	Graut				Petty		
	標本A	B	C	全体	標本A	B	C
平均値	50.1 語	45.5 語	46.9 語	47.5 語	66.1 語	60.2 語	56.3 語
中央値	45.2	38.0	37.4	40.1	56.9	51.3	44.0
四分位範囲	32.1	31.7	39.2	35.5	47.1	44.3	44.7

統計量を用いておこなっている [19].

"The Imitation of Christ" に関しては，その著者の可能性のある人物として Thomas à Kempis と Jean Charlier de Gerson の二人をあげ，彼等の作品と "The Imitation of Christ" の文の長さの平均値，中央値，四分位範囲等を計算し，表1のような結果を得ている．これらの数値の比較から，彼は，à Kempis が著者であると結論している．

一方 "Natural and Political Observations…" は，統計学者 John Graunt の著作とされているが，Yule はこの作品を Graunt の友人で "A Treatise of Taxes and Contributions (租税負納論)" 等で知られる統計学者 Sir William Petty の著作と比較し，表2で示されるようにこの2人の文の長さの分布は異なっていることを明らかにした．

この Yule の研究と同じ様な研究に 1957 年，W. C. Wake がおこなった研究がある [20]. 彼は偽作説がある プラトンの "第七書簡" の文の長さをプラトンの著作「法律」と比べ（表3），その偽作説を否定している．彼はまたこの論文の中で，アリストテレスの著作についても言及している．

表 3. 「第七書簡」と「法律」の比較

	「第七書簡」	「法律」の平均
平均値	28.1 語	27.7 語
中央値	22.3	23.1
第一四分位数	13.1	13.5
第三四分位数	36.0	36.7

C. 特別な指標による推定

ところで，文の長さを用いた推定に満足できなかった Yule は 1944 年，作家の語彙量を測る指標として，K-characteristic と呼ばれる指標を提案し，再度 "The Imitation of Christ" の著者推定をおこなっている[21]. いま，ある作品の中に n_i 回使用された単語が f_i 個あるとすると，この時 $S_1=\sum f_i n_i$, $S_2=\sum f_i n_i^2$ として K は次のように定義される．

$$K = 10^4\left(\frac{S_2}{S_1^2}-\frac{1}{S_1}\right)^{*}$$

彼はある単語が出現するプロセスのモデルとしてポアソン分布を想定した．ポアソン分布では 平均=分散 という関係が成立するが，現実にポアソン分布に従うと考えられる事象を観察すると多くの場合 分散≧平均 となっていること，また，分散を平均で除した変動係数が標本の大きさに影響されないこと等の考察にもとづき，この指標を導出している**．問題の書

* 10^4 は K 値が小さくなりすぎるのを防止するためで他に意味はない．

** この指標の性質については [22], [23], [24], [25] 等で議論されている．

表 4. K 値の比較

作品		K 値
The Imitation of Christ		84.2
à Kempis	1	113.7
〃	2	110.9
〃	3	100.1
〃	4	66.9
〃	5	59.7
Gerson		35.9

"The Imitation of Christ" と à Kempis, Gerson の作品の K 値は表4の様になり，この結果から彼は à Kempis の方が著者の可能性が高いとした．

また A. Ellegård は1962年に発表した論文[26]の中で，1768年から1772年にかけて英国内閣の政策を批判した一連の公開状を"Junius"という偽名で発表した人物が誰であるかを，Distinctiveness ratio という指標を用いて推定している．彼は Junius letter のような18世紀中頃の型にはまった文章に対しては，Yule の K-Characteristic は効力が無いとして，次のような Distinctiveness ratio という指標を提案した．

$$\text{Distinctiveness ratio} = \frac{\text{Junius letter の中でのある単語の相対度数}}{\text{Junius letter 以外の百万語の中のある単語の相対度数}} \quad *$$

彼はこの指標を用いて，100名にもあまる候補者の中から，Junius として Sir Philip Francis の名をあげている．

この2つの指標以外に，語彙量に注目した指標として E. H. Simpson の Index of diversity [27]，K-Characteristic の変形である G. Herdan の K^*-Characteristic [28] などが，また品詞の相対度数に注目した指標として，V. Neubaver & A. Schlisman の動詞一形容詞比 [29]，C. B. Williams の名詞一形容詞比 [30] などが提案され議論されているが，具体的な著者推定問題には用いられていないようであるのでここでは割愛する．

D. χ^2 検定，t 検定による推定

1861年にニューオリンズ・ディリー・クレセント誌に Quintus Curtius Snodgrass という偽名で南北戦争に関する手紙 (Q. C. S. letter) が掲載された．この手紙の著者は「トム・ソーヤの冒険」等の作者 Mark Twain であると一般に信じられてきたが，C. S. Brinegar は1963年 J. A. S. A. に発表した論文 [31] の中で，単語の長さとその頻度に注目して χ^2 検定，t 検定をおこない，この説を否定した．彼は Q. C. S. letter の前の時期，同じ時期，後の時期の Twain の文章から，それぞれ1885語，6106語，2974語の単語を調べ（表5），それをコントロールとして χ^2 検定をおこない，0.5% 水準でも有意な差があることを示した．また，彼は2, 3, 4文字からなる単語の頻度の各々の平均値に関しても t 検定をおこない，それらが5%，1%，5%で有意であることも示している．

1960年に安本は源氏物語の宇治十帖と呼ばれる後半の10巻が前半の44巻と同様紫式部の手になるものであるかを調べた研究を発表している [32], [33]．彼は，源氏物語各巻の (1) 長編度（頁数） (2) 和歌の使用度 (3) 直喩の使用度 (4) 声喩の使用度 (5) 心理描写の数 (6) 文の長さ (7) 色彩語の使用度 (8) 名詞の使用度 (9) 用言の使用度 (10) 助詞の使用度 (11) 助動詞の使用度 (12) 品詞数の12項目について各巻に順位をつけ，U 検定法によって後半の10巻と前半の44巻の差が偶然によって生じる確率を計算し，さらに χ^2 検定をおこなって，宇治十帖と前半の44巻の著者が異なっていると主張している．

また，A. Q. Morton は1965年に新約聖書の中の14通のパウロの書簡** の著者推定に関す

* たとえば uniform (形容詞) の場合には distinctiveness ratio = 0.000280/0.000065 = 4.3 であった．

** ローマ人への手紙，コリント人への第1の手紙，第2の手紙，ガラテヤ人への手紙，エペソ人への手紙，ピリピ人への手紙，コロサイ人への手紙，テサロニケ人への第1の手紙，第2の手紙，テモテへの第1の手紙，第2の手紙，テトスへの手紙，ピレモンへの手紙，ヘブル人への手紙の14通で，いずれも「使徒パウロから…」というような書き出しになっているのでパウロの書簡とされている．

表 5. Mark Twain の著作と Q.C.S. letter の比較

語の長さ	1858年, 1861年の著作	1863年の著作	1867年の著作	Q.C.S. letter
1	74	312	116	424
2	349	1146	496	2685
3	456	1394	673	2752
4	374	1177	565	2302
5	212	661	381	1431
6	127	442	249	992
7	107	367	185	896
8	84	231	125	638
9	45	181	94	465
10	27	109	51	276
11	13	50	23	152
12	8	24	8	101
13以上	9	12	8	61
	1885	6106	2974	13175

る研究を発表している [34]. 彼はパウロの書簡の中に出てくる kai (=and), de (=but), en (=in), autos (代名詞) の頻度を調べ χ^2 検定をおこなって, 次のように結論している.

「パウロの書簡といわれているものの中で, もしガラテヤ人への手紙を書いたのがパウロであるならば, ローマ人への手紙, コリント人への第1の手紙, 第2の手紙もまた彼の手によるものであるといえる. しかしながら, 残りの10通は少なくとも6人の手によるものと考えられる」

E. 判別関数による推定

ハーバード大学の F. Mosteller とシカゴ大学の D.L. Wallace は, 1963年の J.A.S.A. に Federalist paper の著者推定に判別関数を用いた論文を載せている [35]. Federalist paper というのは, 米国憲法の批准を勧めるために 1787 年から 1788 年の間に, Alexander Hamilton, John Jay, James Madison の3人がニューヨーク州の新聞に載せた77編のキャンペーン記事のことで, Hamilton 43編, Jay 5編, Madison 14編, Hamilton & Madison 3編というように12編を除いてはその帰属は明らかにされており, 著者不明の12編も Jay の可能性はなく, Hamilton か Madison のいずれかが著者であるとされていた.

Mosteller はすでに1941年に文の長さによってこの12編の帰属を明らかにしようと試みていたが, その時は表6のような結果を得て失敗に終っていた. そこで1963年の論文では, 彼等は, 記事の内容と無関係で, かつ短い記事の中にでも比較的多く用いられている言葉——たとえば, of, all, also のような言葉——に注目した. その結果, たとえば upon は Hamilton の文章では 3/1000 の割合で使用されているが, Madison の文章では 1/6000 の割合であり, また, to は Hamilton の方が使用率が高く, 反対に by は Madison の方が高い等のことがわかった. そこでこのような識別力の高い31語を用いて

表 6. Hamilton と Madison の文の長さの比較

文の長さ	Hamilton	Madison
平均	34.55 語	34.59 語
標準偏差	19.2	20.3

$$y = \sum_{i=1}^{31} W_i x_i$$

という線型判別関数を作り，これによる判別を試みている．ここで，x_i は帰属不明の記事の中での i 番目の語の出現率で，W_i は i 番目の語の重みであり，この値は Hamilton の記事の場合には y が大きく，Madison の記事の場合には y が小さくなるように次の式で求めている．

$$W_i = \frac{\bar{x}_{i(H)} - \bar{x}_{i(M)}}{R_{i(H)} - R_{i(M)}}$$

ただし，$\bar{x}_{i(H)}$，$\bar{x}_{i(M)}$，$R_{i(H)}$，$R_{i(M)}$ は各々 Hamilton，Madison の文章における i 番目の語の頻度の平均値とレンジである．この判別関数を用いて帰属不明の 12 編の記事の y 値を計算した結果が表7である．次に，彼等2人の著作25編から求めた y 値の平均 $\bar{y}_H$，$\bar{y}_M$，標準偏差 S_H，S_M を用いて t 検定をおこない，No. 55 の記事は疑問が残るもののそれ以外は Madison が著者であると推定している．またこの論文の中で Mosteller 達は確率比を用いての判別も試みている．いま Hamilton，Madison の文章において i 番目の語が出現する確率が，それぞれパラメータ $\theta_{i(H)}$，$\theta_{i(M)}$ のある分布（Mosteller 等はこの分布のモデルとして，ポアソン分布と負の二項分布の2つを想定している）に従っている時に，i 番目の語が x_i 回観測される確率を $f(x_i|\theta_{i(H)})$，$f(x_i|\theta_{i(M)})$ とする．この時

表 7. Hamilton，Madison の記事および著者不明の記事における判別関数値

Hamilon		Madison		著者不明	
記事	y	記事	y	記事	y
13	0.88	134	0.05	49	−0.46
15	0.85	135	−0.50	50	−0.26
16	1.11	201	−0.22	51	−0.76
17	1.26	202	−0.28	52	−0.23
21	1.11	203	−0.27	53	−0.48
22	0.96	204	−0.79	54	−0.46
23	0.62	205	−0.29	55	0.32
25	0.31	206	−0.64	56	−0.01
27	0.80	207	−0.86	57	−0.50
28	0.70	208	−0.38	58	−0.33
29	1.39	209	−0.65	62	−0.41
30	1.32	210	−0.65	63	−0.13
31	1.38	216	−0.16		
32	0.49	217	−0.42		
33	0.75	218	−0.42		
34	1.06	219	−0.53		
35	0.80	220	−0.19		
60	0.73	301	−0.07		
61	0.64	302	−0.22		
67	0.63	311	−0.44		
68	0.34	312	−0.08		
69	1.18	313	0.15		
73	1.19	314	−0.38		
75	1.31	315	−0.59		
76	1.17	316	−0.57		
$\bar{y}$	0.92	$\bar{y}$	−0.38	$\bar{y}$	−0.31
S	0.32	S	0.25		

$$\Pi \frac{f(x_i|\theta_{i(H)})}{f(x_i|\theta_{i(M)})}$$

が大ならば Hamilton が，小ならば Madison が著者の可能性が高いと考えられる．この式を用いた結果やはり Madison が著者であると推定している．

この Mosteller & Wallace の方法を用いて，韮沢は由良物語の著者として加茂真淵説よりも，建部綾足説を支持する研究を発表している [36]，[37]．また，確率比を用いて無量寿経類漢訳者の推定をした研究も報告されている [38]．

F.　因子分析法による推定

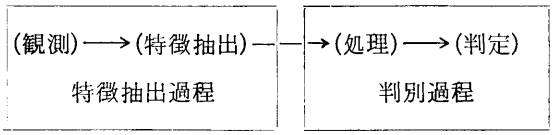
安本は，源氏物語宇治十帖の著者推定を因子分析法によってもおこなっている [39]．彼は前述の 12 項目間の相関係数を計算し，因子分析によって　(1)「比喩多用型—比喩節用型」の因子　(2)「歌物語型—作り物語型」の因子　(3)「体言型—用言型」の因子とも名づけられる 3 つの因子を取り出した．彼はこの因子分析の結果をもとに χ^2 検定をおこない，前の研究と同様，源氏物語の宇治十帖と呼ばれる後半の 10 巻の文体は，前半の 44 巻の文体とやや異なっていると報告している．

G.　その他

ここであげた以外に，具体的に用いられてはいないが，シラブル間の距離やエントロピー [40]，[41]，系列相関 [42] 等も検討されている．

4.　著者推定における問題点

この節では，文体もまた一つのパターンであると考えることによって，著者推定問題を統計的パターン認識の問題に還元し，この観点から前節で紹介した著者推定問題を考察し，問題点を整理する．パターン認識は通常，観測，特徴抽出，処理，判定の 4 つの過程をへるとされているが，ここでは，下記のように特徴抽出と判別の二つに大別して議論を進めることにする．



さて，前節の著者推定問題をこの観点から整理すると表 8 のようになる．

次にこれらの研究で明らかにされた問題点を，特徴抽出，判別の二つの観点から整理してみる．

A.　特徴抽出における問題点

Le style c'est l'homme même（文は人なり）と P. Guiraud は述べているが，文体のいかなる面に著者の特徴が現われるかを知ることが，著者推定問題解決の鍵である．そのためには，少なくとも次の 5 つの側面，イ) 文章の構成要素の性質　ロ) 文法構造　ハ) 音韻　ニ) リズム　ホ) 問題特有の性質（例えば，源氏物語における和歌の数のような）の側面から調べるべきであろう．ところで表 8 が示すように，これまでの研究でとりあげられてきたのは，単語の長さ，文の長さ，単語の使用頻度，語彙量など主に イ) に関するものである．これらについては，次のような問題点が指摘されている．まず単語の長さに関しては，単語の定義が困難であり，その上，文の長さと違い単語の長さは著者の自由にならないため特徴が出にくい等のことが，また文の長さに関しては，単語と同様に文の定義の困難さや，文の長さの測定単位を何にすべきかという問題，さらにはまた，記述体か対話体か等の文の種類によって長さが変る可能性があ

表 8. 主な著者推定問題

問題となった文献，著者	抽出した特徴	用いられた統計量・統計的手法	研究者
シェクスピア，ベーコン	語の長さの分布	モード	T. C. Mendenhall (1887年)
The Imitation of Christ	文の長さの分布	平均値，中央値，四分位範囲，etc.	Undy Yule (1939年)
プラトンの第七書簡	〃	〃	W. C. Wake (1957年)
The Imitation of Christ	語彙量	K-Characteristic	Undy Yule (1944年)
Jonius letter	語の使用頻度	Distinctiveness ratio	A. Ellegard (1962年)
Quintus Curtius Snodgrass letter	語の長さの分布	χ^2 検定，t 検定	C. S. Brinegar (1963年)
新約聖書の中のパウロの書簡	語の使用頻度	χ^2 検定	A. Q. Morton (1965年)
源氏物語の宇治十帖	頁数，和歌数等の12項目	U 検定，χ^2 検定	安本 美典 (1960年)
Federalist paper	語の使用頻度	線型判別関数，確率比	Mosteller & Wallace (1963年)
由良物語	〃	〃	韮沢 正 (1965年)
源氏物語の宇治十帖	頁数，和歌数等の12項目	因子分析法	安本 美典 (1977年)

ることなどが指摘されている．単語の使用頻度に関しては，内容によっては同じ語が多数用いられる可能性があり，したがってどの単語について調べるべきかの決定の困難さが指摘されている．([20], [35] では，文章の内容と無関係で，しかも使用頻度の多い語を用いているが，その選択のために予備調査をおこなっている．)

さらに，語彙量の測度である K-characteristic については [21], [30], [43] 等の結果をみると(データが少なく，明言できないが) 同一作家の作品でも K 値のバラツキが大きいような感じを受け，この測度の使用に関してはさらに研究が必要であろう．

このようにイ) に関しては，問題点はかなり明らかにされているが，ロ)，ハ)，ニ)，ホ) に関しては今後の研究を待たねばならない．これとは別に，Yardi の研究 [14] からもわかるように，文体が著者の一生を通じて不変のものではなく，時とともに変化する可能性があることも常に考慮に入れておかねばならない．この他に，サンプリングおよび測定法の問題点についても [21], [44], [45] 等で議論されている．また具体的な作業をおこなう場合に，解析に用いる資料として何を用いるべきか，また古典の著者推定問題で真筆が存在しない場合にどの写本を用いるべきかというような問題もある (日本の古典の場合に，文の長さが校訂者によって異なる場合があるという報告がある [46])．

B. 判別における問題点

ところで現在までになされた研究は，主にパラメトリックな統計手法によって解析をおこなっている．パラメトリックな方法を用いる際には当然分布型など初めとして種々の統計上の仮定が必要であり，したがって常にそれらの仮定の妥当性が問題となる．著者推定の場合においても，このことは例外ではない．たとえば，語の長さの分布として，対数正規分布 [47]，ワーリング分布 [48]，複合ポアソン分布 [21] などが，文の長さの分布として対数正規分布 [50]，ガンマ分布 [51] が，単語の使用頻度の分布として，ポアソン分布，負の二項分布 [35]，対数指数分布 [49] などが提案されているが，やはりこれらの分布の妥当性についてはまだ検討の余地がありそうである．またこれ以外に，各変数の独立性の仮定など，個々の問題で前提とした種々

の統計上の仮定についてもより詳細な吟味が必要である．勿論このような問題点を避けるために，ノンパラメトリックな方法で判別を試みた例もないではない [32]．しかし，この場合にも各変数の独立性の仮定はおいている．

次に問題になるのは，解析（判別）結果から導びいた結論についてであるが，この点における問題点は実は著者推定問題のタイプと関係させて考えねばならない．一般に著者推定問題は次の2つのタイプに大別できると考えられる．第1のタイプは，問題となっている作品 x が A の著書であるか B の著書であるか，つまり，$x\in A$ か $x\in B$ かの判別問題であり，第2のタイプは x が A の著書であるか否か，つまり，$x\notin A$ かどうかの判別問題である．第1のタイプの著者推定問題としては [21]，[35] などがあげられるが，この場合には，判別関数や，t 検定やあるいは何らかの意味で A や B からの近さを求めてそれらの値を基にして判別がおこなわれている．この場合にも，たとえば線型の判別関数で充分であるか等の統計的方法に関する問題は勿論あるが，しかしより本質的な問題点は x の作者の可能性のある人物をどのようにして A，B の2人に絞るかという統計以前の点にある．たとえば，Yule は [19]，[21] の "The Imitation of Christ" の著者推定において，著者の可能性のある人物として à Kempis と Gerson を考えているが，この2人以外に à Kempis が属していた教団の創立者，ヘーラルト・ホロートが書いたものを à Kempis が修正加筆しながら編集したという説もある*．前述の2人と比較した場合に，ヘーラルト・ホロートの方の可能性が高くなることも考えられないではない．Federalist paper の著者推定 [35] のように著者の可能性のある人物が2人でそれ以外に考えられないというような問題はむしろ著者推定問題としては特殊であって，一般には可能性のある人物を何らかの形で2人に絞って問題を解いている．したがって，この場合には，A と B ではどちらの方がより可能性があるかという結論しか得られない．

一方，2番目のタイプの著者推定問題としては [20]，[32, 33] などがあり，主に t 検定や χ^2 検定等の統計的仮説検定法によって結論を導出している．この場合に，仮説が棄却されたなら $x\notin A$ と結論でき問題はないが，仮説が棄却されなかった場合は注意しなければならない．仮説が棄却されなかったことが即 $x\in A$ とはならないからである．この意味で [20] の結論導出には問題がある．

5. おわりに

この稿でとり上げた研究例の多くは欧米のものである．したがって，由良物語の著者推定のように，Mosteller 達が Federalist paper に対して用いたのと同じ手法である程度の成果をあげた例はあるにしろ，しかし一般的に，言語体系が異なる日本文の場合に，これらの研究成果がどの程度役立つかは，今後の研究を待たねばならない．また，前節で指摘したように，統計的手法による判定は客観性を持つという反面，多くの仮定のもとでの結論であることに留意する必要がある．したがって，最終的な結論は，統計的な研究も含め多くの側面からの研究を通じてなされるべきである．

以上，著者推定問題と呼ばれるこの種の問題に対する今後の研究の一つの道標となればと願い，この分野の研究の展望をおこなったが，著者の知識の不足から，重要な研究を欠落していたり，引用論文の解釈に多くの誤謬があるかもしれずこの点に関し読者の御叱正を乞う次第である．

なお，本稿をまとめるにあたり，レフリーの方々より多くの有益な助言をいただきました．ここに感謝の意を表します．またこの研究は文部省科学研究費一般研究 B「統計的決定理論の立場からの文献学的判別問題に対する研究」の一部としておこなわれたことを付記しておきます．

* 安本 [39] 参照

参 考 文 献

[1] Campbell, L. (1867). The Sophistes and Politicus of Plato, Oxford, Clarendon Press.
[2] Dittenberger, W. (1896). Sprachliche Kriterien für die Chronologie der Platonischen Dialoge, Hermes, 16.
[3] Lutosławski, W. (1897). The Origin and Growth of Plato's Logic, with an Account of Plato's style and the Chronology of his Writings.
[4] Ritter, C. (1935). Unterabteilungen innerhalb der zeitlich ersten Gruppe platonischen Schriften, Hermes, 70.
[5] Cox, D.R. and Brandwood, L. (1959). On a Discriminatory Problem Connected with the Works of Plato, J.R. Statist. Soc., B. 21, 195–200.
[6] Bathurst, C. (1857). Remarks on the Difference in Shakespeare's Versification in different Periods of his Life, London, Parker and Son.
[7] Ingram, J.K. (1874). On the weak endings of Shakespeare, New Shakes. Soc., Trans., Ser. 1, Part 3, 442–446.
[8] Fleay, F.G. (1874). On metrical tests as applied to dramatic poetry. I. Shakespeare, New Shakes. Soc., Trans., Ser. 1, Part 1, 1–16; Postscript, 38–39; also Discussion, 17–37.
[9] Furnival, F.J. (1887). Introduction to the Leopold Shakspeare, London, Cassell & Co.
[10] Furnival, F.J. (1907). "Loues Labors Lost": the Old-Spelling Shakespeare. London, Chatto and Windus.
[11] König, G. (1888). Der Vers in Shakspeares Dramen.
[12] Conrad, H. von (formerly Isaac) (1897). Shakesperes Selbstbekenntnis, Stuttgart.
[13] Chambers, E.K. (1930). William Shakespeare: A Study of the Facts and Problems, Oxford, Clarendon Press, 2 vols.
[14] Yardi, M.R. (1946). A statistical approach to the problem of the chronology of Shakespeare's plays, Sankhyā, 7(3), 263–268.
[15] Lord, R.D. (1958). De Morgan and the statistical study of literary style, Biometrika, 45, 282.
[16] de Morgan, Sophia (1882). Memoir of Augustus de Morgan, by his wife Sophia Elisabeth de Morgan, with Selections from his Letter, London.
[17] Mendenhall, T.C. (1887). The characteristics curves of composition, Science, 11, 237–249; Supplement, March 1887.
[18] Mendenhall, T.C. (1901). A mechanical solution of a literary problem, Popular Science Monthly, 60(7), 97–105.
[19] Yule, G.U. (1939). On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to two Cases of Disputed Authorship, Biometrika, 30.
[20] Wake, W.C. (1957). Sentence-Length Distributions of Greek Authors, J.R. Statist. Soc., A, 20, 331–346.
[21] Yule, G. Undy (1944). The Statistical Study of Literary Vocabulary, Cambridge University Press.
[22] Williams, C.B. (1946). Yule's "Characteristic" and the Index of Diversity, Nature, 157.
[23] Good, I.J. (1953). The population frequencies of species and the estimation of population parameters, Biometrika, 40.
[24] Herdan, G. (1955). A new derivation and interpretation of Yule's "Characteristic K", J. of Applied Mathematics and Physics, VI, 332–334.
[25] Herdan, G. (1958). The mathematical relation between Greenberg's index of linguistic diversity and Yule's characteristic, Biometrika, 45.
[26] Ellegård, A. (1962). A Statistical Method for Determining Authorship: The Junius Letter, 1769～1772, Gothenburg Studies in English, no. 13.
[27] Simpson, E.H. (1949). The measurement of diversity, Nature, 163, 688.
[28] Herdan, G. (1966). The Advanced Theory of Language as Choice and Chance, New York.
[29] Antosch, F. (1969). The diagnosis of literary style with the verb-adjective ratio, Statistics and Style, 57–65.
[30] Williams, C.B. (1970). Style and Vocabulary, Hafner Publishing Co., New York.
[31] Brinegar, C. (1963). Mark Twain and the Quintus Curtius Snodgrass Letters: a statistical test of authorship, J. Amer. Statist. Ass., 58, 85–96.
[32] 安本美典 (1957). 宇治十帖の作者—文章心理学による作者推定—, 文学・語学, 1957, 4.

[33] 安本美典 (1958). 文体統計による筆者推定 —源氏物語, 宇治十帖の作者について—, 心理学評論 Vol. 2, No. 1.
[34] Morton, A.Q. (1965). The authorship of Greek prose, J.R. Statist. Soc., A, 128 (2), 169–233.
[35] Mosteller, F. and Wallace, D.L. (1964). Inference in an Authorship Problem, J.A.S.A., 58, 275-309.
[36] 韮沢 正 (1965). 由良物語の作者の統計的判別, 計量国語学 No. 33.
[37] 韮沢 正 (1973). ある統計的判別の試み, 数理科学 No. 117, 3月号.
[38] 後藤義乗 (1978). 数理文献学的方法による無量寿経類漢訳者の推定, 印度学仏教学研究第26巻第2号.
[39] 安本美典 (1977). 現代の文体研究, 岩波講座日本語, 397-423.
[40] Fucks, W. (1954). On mathematical analysis of style, Biometrika, 39.
[41] Fucks, W. (1954). On nahordnung and fernordnung in samples of literary texts, Biometrika, 41.
[42] 米田正人 (1974). 文の長さの統計学的一考察, 電子計算機による国語研究 VII, 国立国語研究所.
[43] Bennett, P.E. (1957). The statistical measurement of a stylistic trait in Julius Caesar and As You Like It, Shakespeare Quarterly, VIII, 33–50.
[44] 安本美典 (1957). 文の長さの分布型, 計量国語学 No. 1.
[45] 林知己夫 (1957). 安本氏の「文の長さの分布型」について, 計量国語学 No. 2.
[46] 三上悠紀夫 (1959). 古典校訂本の句点は信用できるか, 計量国語学 No. 9.
[47] Williams, C.B. (1956). A note on an early statistical study of literary style, Biometrika, 45, 248–256.
[48] Herdan, G. (1961). Quantitative Linguistics, London, Butlerworth.
[49] 安本美典 (1977). 語彙の量的構造, 数理科学 No. 168, 6月号.
[50] Williams, C.B. (1940). A note on the statistical analysis of sentence length as a criterion of literary style, Biometrika, 31, 356–361.
[51] 佐々木和枝 (1976). 文の長さの分布型, 計量国語学 No. 78.