補足資料１

よくある質問へのヒント

・構成要素，異なり構成要素の分布の特性
・有意性テスト（とくに頻度による有意性テスト）

大隅　昇
テキスト・マイニング研究会代表
統計数理研究所・名誉教授

ここでは**構成要素，異なり構成要素とその分布**の観察および**有意性テスト**（とくに頻度による有意性テスト）に関連した事項について要約する．なお WordMiner が提供する情報の理解と解釈には，若干の統計的データ解析の知識を必要とし，よってここに記述する情報も，より正確に理解するためには初等的な統計学の知識があることが望ましい．

1． 構成要素の観察の要領

テキスト型データの解析時の重要かつ基本的操作として「**構成要素，異なり構成要素**」の分布の観察がある．とくに、「異なり構成要素変数」から基本情報を観察して，以後の分析（とくに多次元データ解析）への入り口となるデータ表を確定する操作が大切である．

1．1 構成要素数の分布の観察

一般に，単語・語句・語彙の出現頻度を計量的に分析評価する方法として，ジフの法則（Zipf's law）やパレートの法則（Pareto's law）などが知られている．これについては 1．6 節で簡単に述べる．

ここではまず，WordMiner で得られる構成要素，異なり構成要素の分布の観察方法について要約する．WordMiner では，構成要素数（累積構成要素数，閾値別の構成要素，そしてそれぞれに対する異なり構成要素数）を「構成要素数の情報１」「構成要素数の情報２」として算出表示する（図 1，図 2）．

これらの基本的な意味解釈は既にセミナーで確認したところであるから，ここでは，これらの情報から得られる，より一般的な関係について述べる．

構成要素変数【[V0227] Q3_1:あなた自身にとって「インターネット」は一分かち書き一編集_all】/構成要素変数の情報

	閾値の水準	その頻度以上の総構成要素数	その頻度以上の異なり構成要素数	その頻度以上の異なり構成要素率
1	頻度 1	8609	2242	26.0
2	頻度 2	7145	778	10.9
3	頻度 3	6569	490	7.5
4	頻度 4	6185	362	5.9
5	頻度 5	5917	295	5.0
6	頻度 6	5702	252	4.4
7	頻度 7	5546	226	4.1
8	頻度 8	5350	198	3.7
9	頻度 9	5174	176	3.4
10	頻度 10	4976	154	3.1
11	頻度 11	4826	139	2.9
12	頻度 12	4749	132	2.8
13	頻度 13	4617	121	2.6
14	頻度 14	4513	113	2.5
15	頻度 15	4443	108	2.4
16	頻度 16	4323	100	2.3

構成要素変数の情報1 | 構成要素変数の情報2 | 置換の結果 | 削除の結果

図 1「構成要素変数の情報１」の内容

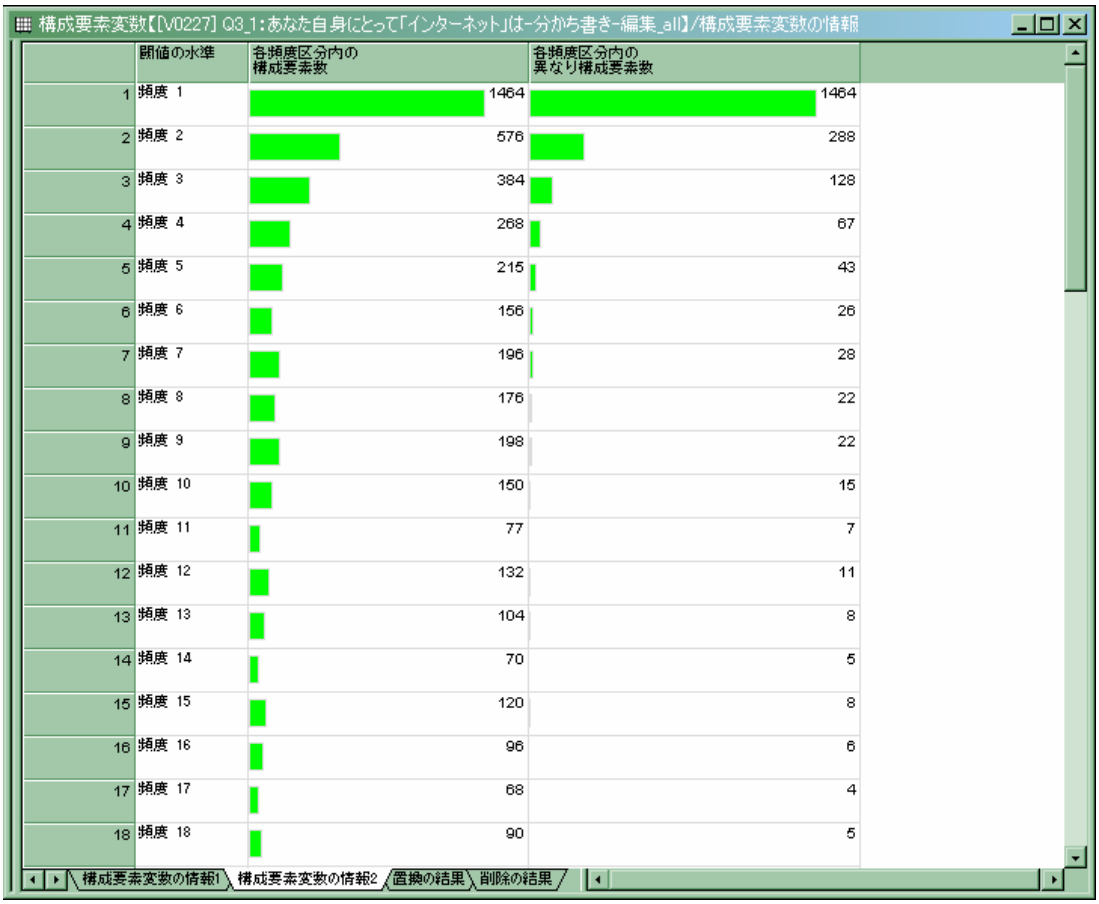

図 2「構成要素変数の情報 2」の内容

1．2　　構成要素数と異なり構成要素数の関係　－実験調査結果を用いて－

過去に行ったいくつかの実験調査（意識調査，市場調査など）で得られた知見から抜粋引用した例をいくつか示す．

例１：

自由回答質問で，回答記入形式を変えた場合の結果の差異を調べた．ここで調査方式（調査モード）としては郵送法，Web 調査法（いずれも自記式）を用いた．質問と用いた自由回答の記入形式を表 1 に要約してある．

表 1　用いた質問と回答記入形式

質問の内容	回答記入形式
Q1-1：ブランド名の印象	テキスト・ボックス（A）
Q1-2：商品／事業想起	連記型（5 項目）
Q1-3：ふさわしい言葉（形容詞などを用いて挙げる）	連記型（10 項目）
Q6：企業の長所	テキスト・ボックス（B） （*）ボックス並置／一部イメージ貼付
Q7: 自宅のパソコンの購入目的と購入の経緯	テキスト・ボックス（A）
Q3-8：自宅のパソコンの使用状況	テキスト・ボックス（A）
Q6：パソコンの達人のイメージ	テキスト・ボックス（A）

ここで得られた構成要素数，異なり構成要素数（率）の分布が図 3，4 である．これらを観察すると，以下の特徴がある（要約）．

① 調査方式（Web 調査と郵送調査）の違いは見かけ上はあまりみられない（図中にマーカーは入れていないがそういう傾向があった）．
② テキスト・ボックスと連記型の違いは明らかにある．また，テキスト・ボックスのバラツキが大きい．テキスト・ボックスの作り方（大きさ，スクロール・バーの有無など）が影響するものと思われる．
③ 書込領域を限定した連記型（5 項目），連記型（10 項目），大きさの異なるテキスト・ボックスに対応して書き込み量が増える．これも予想される特徴である．
④ 総構成要素数（総単語数）が増えると異なり構成要素数（異なり単語数）も増えるが次第に伸び方が低減する（指数的に変化する）．（サンプル数が同じ状況で）書き込み量・記入量が増えると，それは頭打ちとなると傾向にある．この現象はある程度予想されることで，書き込み量が増えても，ある設問についての回答内容はいずれある量に収まることを示している．これは 1.6 節に記すジフの法則などにも関連する．

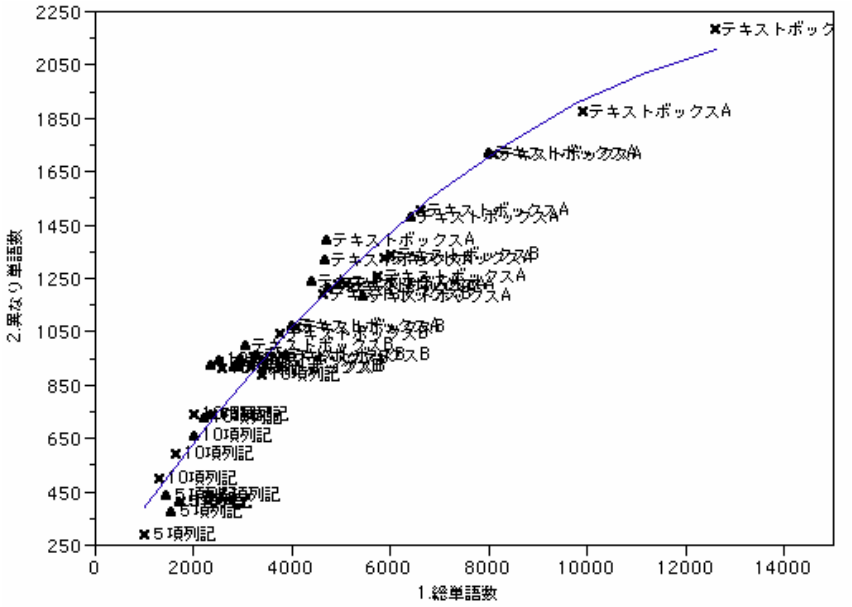

図 3　構成要素数と異なり構成要素数の関係

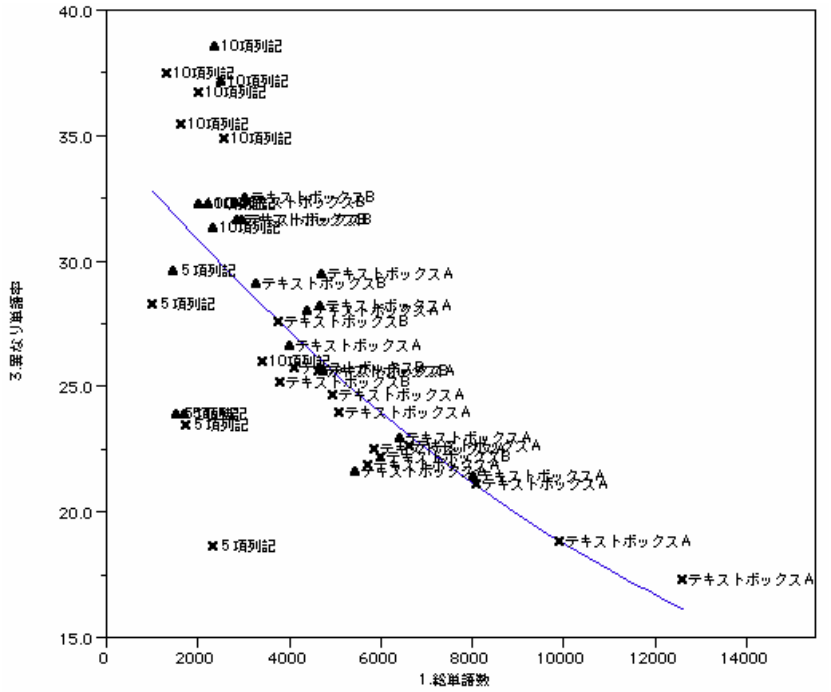

図 4　総構成要素数と異なり構成要素率の関係

例２：

次の例として意識調査で得た結果をみる．ここでは，調査方式として主に Web 調査を用いているが，一部に郵送調査が含まれる．いずれも「自記式」という点では共通している．

得られた構成要素数，異なる構成要素数（率）の関係をグラフとした（図 5，6）．図 5 は例 1 と同様に横軸が構成要素数，縦軸が異なり構成要素数を示したものである．調査方式や調査票の質問形式等に違いがあることに注意して観察する必要があるが，いくつかの特徴がみられる．

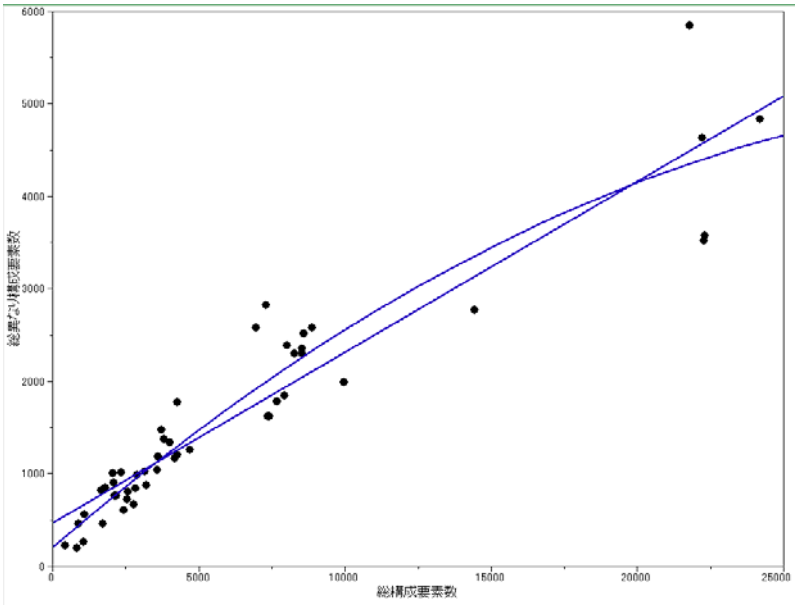

図 5　構成要素数と異なり構成要素率の関係

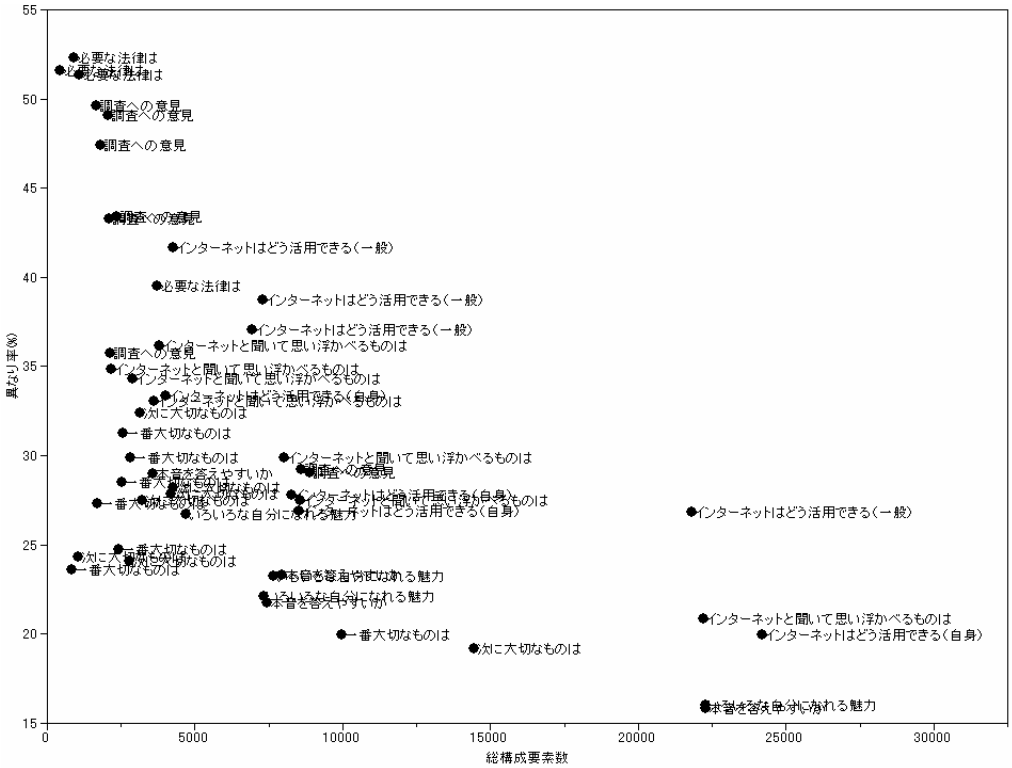

図 6　構成要素数と異なり構成要素率の関係（ラベル付）

例 1 でもそうであったが，一般に構成要素数の増加に伴い，異なり構成要素数は指数的に増加するが単純な比例関係では増えることはない（断定的には言えないが概ねそのような傾向を示す）．他の実験調査でも類似の結果を得ている．とくに，同じ調査環境（調査方式，調査票などをほぼ同じ条件に揃えて行うなど）で実施した意識調査では，質問内容（ワーディング），回答記入形式（列記型，テキスト・ボックスの大きさなど）によって，構成要素数の出現頻度に違いがあることが分かっている．

ここで図 5 のデータにつき縦軸を異なり構成要素率（%）とし，さらに用いた質問文の一部をラベルとして付与してみると，別の情報が見えてくる（図 6）．図 5 の主な傾向としては，構成要素数の増加に伴い異なり構成要素率が低減するように見えたが，図 6 として質問内容の見出しを見ると（やや見にくいが），実は質問内容によってもかなり分かれていることが見える．

別の特徴として，質問文の意図する内容（何を聞きたいのか）やワーディングと構成要素数などの得られる計量的な情報との関係にも注意すべきである．例えば，上の例に登場した質問であると以下の傾向がみられる．

①回答文の中の書込量が少なく，また現れる単語・語句が特徴的であるような質問文の例

質問文例 1：

「あなたにとって，一番大切だと思うものは何ですか．１つだけあげてください．」

「この他に，大切だと思うものとして何がありますか．いくつでもあげてください．」

質問文例 2：

「あなたの好きな食べ物は何ですか．」

②回答の書込量が比較的多くなる傾向にある質問文の例

質問文例 3：

「あなたご自身にとって「インターネット」は，どのようなことがらに活用できると思いますか．どんなことでも結構ですので，以下になるべく具体的にご記入ください．」

質問文例 4：

「では，一般的に「インターネット」は，どのようなことがらに活用できると思いますか．なるべく，他にはないような活用法を，どんなことでも結構ですので，以下になるべく具体的にご記入ください．」

ここで，「一つだけ…」「いくつでも…」「どんなことでも…具体的に…」といったワーディングを使っていることにも注意しよう．この他，通常の社会調査における質問作成上の留意事項にも配慮することはもちろんである（ワーディング，文脈効果，ダブル・バーレル，質問文の配置など）．

１．３　構成要素数，異なり構成要素数（率）の特徴と観察上の留意点

こうした実験結果から（その他の多数の経験から），意識調査などの例では，経験則的には以下のような特徴が観察されている．

① 調査方式（調査モード）による差異がある
- 自記式，面接員記入方式
- 面接法，留置法，等々
- Web 調査，電話調査，郵送調査，…

② 調査票の自由回答記入形式による差異がある
- 箇条書きか列記式か
- テキストボックスのサイズと書込量の制限（文字数）の有無

③ 同一の調査対象者に対して調査方式を変えたとき，必ずしも同じ結果とはならない
④ 文脈効果，調査票内での自由回答質問の置かれる位置の影響があること
⑤ 調査票の質問のボリューム（質問数）の多少が影響すること
⑥ 同じ質問文（自由回答質問）を用いて，もし等質の回答者集団であるとしてサンプル数が増えたとすると，異なり構成要素率が低減すること

なお一般的な傾向として，
- サンプル数（回答数）が少ないときには，（見かけ上は）異なり構成要素率が大きくなる傾向にある
- サンプル数が非常に大きくなると，異なり構成要素率が小さくなる傾向にある

これらの性質は上にみた例のような傾向があることを考えると，当然な特徴のように見える．要するに，構成要素数，異なり構成要素数（率）の分布には，確定的な規則やきまりがあるのではなく(注)，**データ収集方式（data collection mode）に依存して決まる**というきわめて常識的なことが分かってくる．

なお異なり構成要素率の大小は，一見すると「語彙数の潤沢度」を反映しているように見える．また一部の研究にそのような報告も見られるが，上にみたいくつかの例からも明らかなように，このことは一般的な規則とは思われない（とくに，上に示した質問文への回答など）．

（注）ここで言う「確定的な規則やきまりがあるのではない」とは，1.6 節で後述する総構成要素数，異なり構成要素数等に見られるジフの法則のような，ある種の規則性のことを言うのではなく，構成要素の編集や再加工を行うための目安はない，という意味である．

1．4 さらに一般的な留意事項

以上のことから，“分析に用いる構成要素数や異なり構成要素数（率）”の解釈には，こうすべきであるという規則や法則があるのではなく，自分が意図するようなデータを取得できるような，自由回答・自由記述データ取得上の実験計画，データ取得法設計に依存するということになる．

以上を勘案すると WordMiner における「**閾値の指定**」も確定的な決まりがあるわけではなく，経験則的に以下のような対応を行うことが求められる．

- 社会調査（意識調査，態度調査，市場調査でのアンケートなど）であれば，調査票に含めた複数の自由回答質問について，構成要素数，異なり構成要素数（率）の分布を上の例のように比較分析し質問文の特徴を吟味する．
- グループ･インタビューやフォーカス･グループなどの場合も，同様の観察が必要である．
- エスノグラフィーなどで利用するインタビュー，あるいは日記形式データ，発話分析などでは，1 回答として扱う単位の大きさの決め方と構成要素数，異なり構成要素数の関係を吟味して決める．

つまり，テキスト型データの扱いについては，これといった決まりはなく，以下のように考えることになる．

① ケース・バイ・ケースであると心得ること（テキスト型データの分析の特徴でもある）．
② しかし，構成要素数と異なり構成要素数（率）の観察は不可欠であること．
③ 同一サンプルについて，異なる質問文（自由回答）を複数用意したときには，必ず構成要素数，異なり構成要素数（率）を比較分析すること．
   - 質問によって構成要素数・異なり構成要素数の分布傾向が異なるか否か？
   - 構成要素数が多い質問と少ない質問の特徴はどう違うか？
   - 書き込みが少ない，つまり空白が多い質問はどれか？あるいは逆の特性を示すか？

・ それらの理由として何が考えられるか？
・ 自由回答質問の回答内容（構成要素の分布）の観察から，質問文が意図したように機能したかを知ること

④ 初動探査としては，閾値をかなり大き目に設定して，つまり出現頻度の少ない状況で構成要素の特徴的な単語・語句を観察するとよい．

いずれにしてもこれらはコンピュータや WordMiner を利用するスキルに関わることと言うよりも，調査方法論とくに自由回答質問の設計やインタビュー方式など，テキスト型データ取得方法のリテラシーに関連することである．

1．5 異なり構成要素の処理上の留意事項

一般的に以下のような点に注意して分析処理を進める必要がある．

① 異なり構成要素数の頻度分布は「出現頻度数 1 が最も多く」以下指数的に低減するという特徴があること．

② 指数的に低減する傾向を示すが，出現頻度順位と対応する構成要素数やその頻度との間には，ある種の規則・法則性があることが予想されるが，このことがデータ解析の直接の指針となるかは検討余地がある（1.6 節も参照）．

③ 構成要素選出の閾値としてデフォルト（標準値）を「2」とした理由は，まず出現頻度がもっとも多い「1 度しか登場しない構成要素」を除外するという手当を行い，以後急速に減少する構成要素数の分布の特徴を観察する，つまり初探査的にデータに内在する“主たる”特徴（潜在する構造，規則性）をまず知ることに努めること．

④ 次第に閾値を増やしたときに（つまり構成要素数を減らしたときに），上でみた構造にどのような変化が現れるかを観察する．初動探査的には，むしろ始めに「閾値」を大きく設定し，構成要素群の主要な特徴をしらべる方がよい．

⑤ しかし「１回しか登場しない構成要素に意味がある」との立場で分析をしたいこともあるだろう．この場合，類語語句を纏めて出現頻度を大きくなるようにして復活させるかという方法がある．つまり類語・関連語の要約（シソーラスの扱い）をどう考えるかという課題に関連する．

⑥ そもそも，統計的データ解析の観点からは，ある程度まとまった数のデータの中にある潜在的な特徴を探査する，つまりは探索的にマイニングすることを目標としているのであって，いわゆるレアケースやインシデントな場合の分析にはあまり適していないと心得るべきである（別のアプローチが必要）．

1．6 単語･語句の古典的な計量化の方法

一般に，単語・語句・語彙の頻度に関する法則として，

・ ジフの法則（Zipf’s law）
・ パレートの法則（Pareto’s law）

などが知られている．例えば，大まかに（単語頻度）×（出現頻度の順位）≒（一定数となる）などの法則がある（ジフの法則）．これについて簡単な例を示しておこう．こうした法則の吟味は，単語・語句数がかなり多い場合，例えば，小説・文芸書あるいは日記形式などの文章解析・内容分析などの場合に効果的と思われる．もちろん社会調査などで用いる自由回答質問でも，構成要素数の出現頻度分布のこうした法則の成り立ちを確認するための初動探査的な計量化法として有用である．ここでは，Baayen（2001）による下記の文献を参考に例を用いて示す．

R. Harald Baayen（2001）：*Word Frequency Distributions*, Kluwer Academic Publishers.

まずジフの法則について簡単に触れ，実際の数値例でその特徴・傾向を観察しよう．はじめに上記の Baayen の文献にならって，若干の記号と数式を用意する．

N: 全体の単語数，つまり総構成要素数
（分析対象とするコーパスに相当）
$V(N)$: 総構成要素数Nに占める異なり構成要素数
m: 構成要素の頻度分布における閾値，つまり出現頻度回数
$V(m,N)$: ある出現頻度区分m内の異なり構成要素数
$g(m,N)$: ある頻度m以上（閾値m以上）の異なり構成要素数
$G(m,N)$: ある頻度m以上（閾値m以上）の構成要素数
z: 単語（構成要素）の出現頻度を大きさの順に並べたときのある単語の順位
$f_z(z,N)$: 順位zとなったある単語の出現頻度

以上の約束のもとに，各数量に以下の関係がある．

$$N = \sum_m mV(m,N) \tag{1}$$
$$V(N) = \sum_m V(m,N) \tag{2}$$
$$V(m,N) = g(m,N) - g(m+1,N) \tag{3}$$
$$g(m,N) = \sum_{w \geq m} V(w,N) \tag{4}$$

このとき、いわゆる**ジフの順位頻度分布**（Zipfian rank-frequency distribution）とは以下のような関係をいう（式（5）～（7））．

$$g(m,N) = z \quad \Leftrightarrow \quad f_z(z,N) = m \tag{5}$$

ここで，m は構成要素の頻度分布における閾値（出現頻度回数），$g(m,N)$ はある m 以上の異なり構成要素数である．また z は，ある単語（構成要素）の出現頻度を大きさの順に並べたときのある単語の順位を示し，また $f_z(z,N)$ は順位 z となったある単語の出現頻度である．ジフはこのことから，以下の関係が（経験的に）なり立つのではないかと予想を立てた．

$$f_z(z,N) = \frac{C}{z^a} \tag{6}$$

ここで a はある定数，C も調整用のパラメータである（次の式（7）の回帰係数と切片に相当）．これと上の関係式から，式（6）の両辺の対数（常用対数）をとると以下が示される．

$$\begin{aligned} \log f_z(z,N) &= \log \frac{C}{z^a} \\ &= \log C - a \log z \\ &\Updownarrow \\ \log m &= \log C - a \log g(m,N) \end{aligned} \tag{7}$$

つまり，単語出現頻度とその順位との間には，対数をとると（ほぼ）直線の関係にある．厳密な法則ではないが実際に経験的にはこうした傾向がほとんどのデータで見られる．いま

これをいくつかの例で調べることにしよう．

例１：
まず，Web 調査の次の自由回答質問文から得たデータを調べる．

＜質問文１＞
「あなたにとって，一番大切だと思うものは何ですか．１つだけあげてください．」
「この他に，大切だと思うものとして何がありますか．いくつでもあげてください．」

ここでは，この 2 問を併合し「あなたにとって大切だと思うものは何ですか」として扱う（以下，「大切なもの」と略記）．各数値の関係を表 2，表 3 とした．これらの各数値，とくに m，$V(m,N)$，$g(m,N)$，z，$f_z(z,N)$ それぞれの関係を比較分析することが，一つの計量的評価法として利用できるのである．

表 2　構成要素数の分布：$g(m,N)$，$V(m,N)$ の関係

出現頻度（m） つまり 閾値の水準（m）		$G(m,N)$	$g(m,N)$	$g(m,N)/G(m,N)\times100$	$V(m,N)$
		その頻度以上の総構成要素数	その頻度以上の異なり構成要素数	その頻度以上の異なり構成要素率（%）	各頻度区分内の異なり構成要素数
1	頻度 1	20693	3767	18.2	2366
2	頻度 2	18327	1401	7.6	486
3	頻度 3	17355	915	5.3	228
4	頻度 4	16671	687	4.1	139
5	頻度 5	16115	548	3.4	87
6	頻度 6	15680	461	2.9	58
7	頻度 7	15332	403	2.6	46
8	頻度 8	15010	357	2.4	35
9	頻度 9	14730	322	2.2	38
10	頻度 10	14388	284	2.0	27
11	頻度 11	14118	257	1.8	15
12	頻度 12	13953	242	1.7	26
13	頻度 13	13641	216	1.6	17
14	頻度 14	13420	199	1.5	20
15	頻度 15	13140	179	1.4	9
16	頻度 16	13005	170	1.3	14
17	頻度 17	12781	156	1.2	7
18	頻度 18	12662	149	1.2	10
19	頻度 19	12482	139	1.1	12
20	頻度 20	12254	127	1.0	6
21	頻度 21	12134	121	1.0	4
22	頻度 22	12050	117	1.0	5
23	頻度 23	11940	112	0.9	5
24	頻度 24	11825	107	0.9	5
25	頻度 25	11705	102	0.9	2
26	頻度 26	11655	100	0.9	2
27	頻度 27	11603	98	0.8	5
28	頻度 28	11468	93	0.8	1
30	頻度 30	11440	92	0.8	3
～ 省 略 ～					
388	頻度 388	4136	6	0.1	1
527	頻度 527	3748	5	0.1	1
552	頻度 552	3221	4	0.1	1
661	頻度 661	2669	3	0.1	1
774	頻度 774	2008	2	0.1	1
1234	頻度 1234	1234	1	0.1	1

表 3 構成要素(単語)とその出現頻度分布，他
順位 z と出現頻度 $f_z(z,N)$，それらの対数の関係

順位（z）	構成要素（単語）	順位の対数 $\log z$	構成要素出現頻度 $f_z(z,N)$	$\log f_z(z,N)$	回答者数
1	家族	0.000	1234	3.091	1181
2	自分-自分自身	0.301	774	2.889	612
3	友人-仲間	0.477	661	2.820	631
4	金	0.602	552	2.742	532
5	健康	0.699	527	2.722	497
6	仕事	0.778	388	2.589	361
7	人	0.845	380	2.580	310
8	生活	0.903	313	2.496	292
9	趣味	0.954	298	2.474	291
10	時間	1.000	287	2.458	251
11	ある	1.041	274	2.438	241
12	する	1.079	245	2.389	210
13	できる	1.114	199	2.299	176
14	です	1.146	182	2.260	122
15	親-両親-兄弟	1.176	178	2.250	148
16	して	1.204	176	2.246	159
17	子供-子供達	1.230	173	2.238	155
18	いる	1.255	164	2.215	147
19	思いやり	1.279	141	2.149	136
20	心	1.301	133	2.124	121
21	気持	1.322	127	2.104	116
22	ゆとり-余裕	1.342	115	2.061	109
23	ない	1.362	114	2.057	101
24	大切	1.380	113	2.053	95
25	幸福	1.398	108	2.033	99
26	愛-愛情	1.415	107	2.029	105
27	生きる-生きて	1.431	104	2.017	98
28	家庭	1.447	103	2.013	103
29	した	1.462	94	1.973	92
30	人間関係	1.477	91	1.959	87
～ 省 略 ～					
3754	話し合い	3.574	1	0.000	1
3755	話し相手	3.575	1	0.000	1
3756	話す	3.575	1	0.000	1
3757	惑わされ	3.575	1	0.000	1
3758	惑わされず	3.575	1	0.000	1
3759	枠内	3.575	1	0.000	1
3760	嗅覚	3.575	1	0.000	1
3761	嗜好	3.575	1	0.000	1
3762	穢れて	3.575	1	0.000	1
3763	絆仕事	3.576	1	0.000	1
3764	絆・人間関係	3.576	1	0.000	1
3765	謳歌	3.576	1	0.000	1
3766	飄々	3.576	1	0.000	1
3767	騙したり	3.576	1	0.000	1

とくに表 2，表 3 から，以下の重要な関係を読み取ることができる（式（5）の確認）.

$$
\begin{array}{ccc}
g(m,N)=z & \Leftrightarrow & f_z(z,N)=m \\
\Downarrow & & \Downarrow \\
g(1234,N)=1 & \Leftrightarrow & f_z(1,N)=1234 \\
g(\ 774,N)=2 & \Leftrightarrow & f_z(2,N)=774 \\
\cdots\cdots & & \cdots\cdots \\
\cdots\cdots & & \cdots\cdots \\
g(1,N)=3767 & \Leftrightarrow & f_z(3767,N)=1
\end{array}
$$

さらに，式 (7) の関係を調べるために図を描いてみる．まず横軸を $\log z$，縦軸を $\log f_z(z,N)$ として上のデータを図に表すと図 5 のようになる．

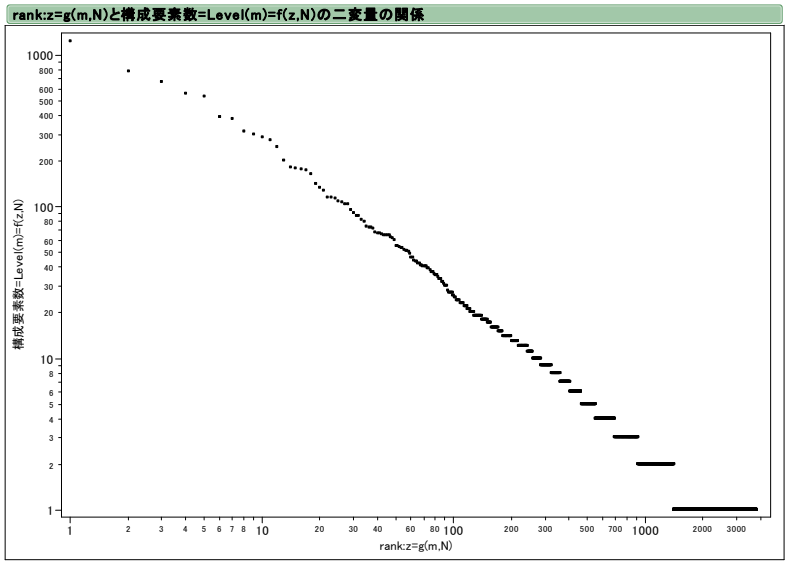

図 5　$\log z$ と $\log f_z(z,N)$ の関係

これを見ると確かに（ほぼ）直線的な関係にある（式（7）が成立しているようだ）．では別の例ではどうであろうか.

例２：

次の 2 つの質問文をとりあげる．用いた調査方式はいずれも Web 調査である.

＜質問文１＞

「あなたにとって，一番大切だと思うものは何ですか．１つだけあげてください.」
「この他に，大切だと思うものとして何がありますか．いくつでもあげてください.」

ここでは前の例と同様，この 2 問を併合して「あなたにとって大切だと思うものは何ですか」として扱う（以下，「大切なもの」と略記）.

＜質問文２＞

「あなたご自身にとって「インターネット」は，どのようなことがらに活用できると思いますか．どんなことでも結構ですので，以下になるべく具体的にご記入ください.」

質問文 1 は，調査サイト T 社，質問文 2 は，2 つの調査サイト（T 社，D 社）で行った調査結果を用いる．それぞれの詳しい数値は挙げないが，以下の 2 種類の統計値のグラフを観察すれば，ジフの法則の意味が理解されるだろう．

①m の対数 $\log m$ と $g(m,N)$ の関係，あるいは m の対数 $\log m$ と $\log g(m,N)$ の関係
②単語（構成要素）の出現頻度の対数 $\log f_z(z,N)$ とその順位 z の対数 $\log z$ の関係

これは，式（5）あるいは式（7）の関係を図で確かめることである．ここでは①について 2 つのグラフを描いてみる．図 6 は m の対数 $\log m$ と $g(m,N)$ の関係を示す図である．明らかに指数的に低減する傾向にある．これを m の対数 $\log m$ と $\log g(m,N)$ の関係として図 7 のように示すと式（7）の予想，つまり両軸対数ではほぼ直線的になるという傾向が顕著に見られる．ただし，用いた質問文，調査サイトの関係などの特徴が何かは直ちには分からない．

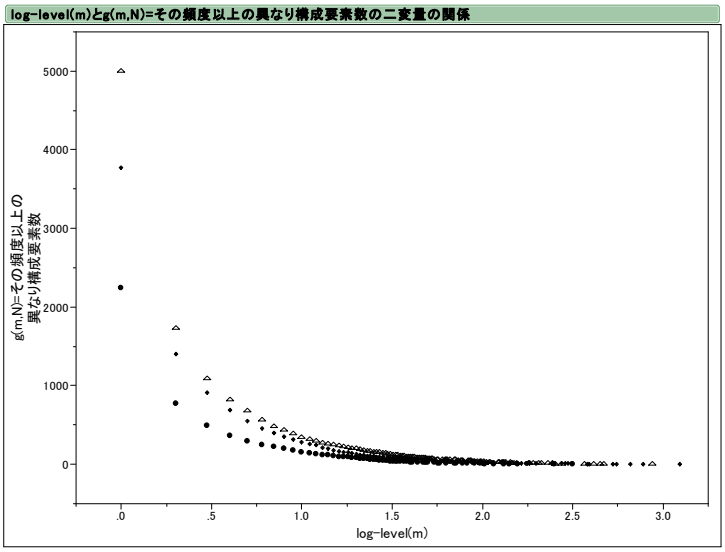

図 6　$\log m$ と $g(m,N)$ の関係

記号の対応：質問 2 の2調査サイト（●＝T 社、△＝D 社），質問 1 の「大切なもの」（◇＝T 社）

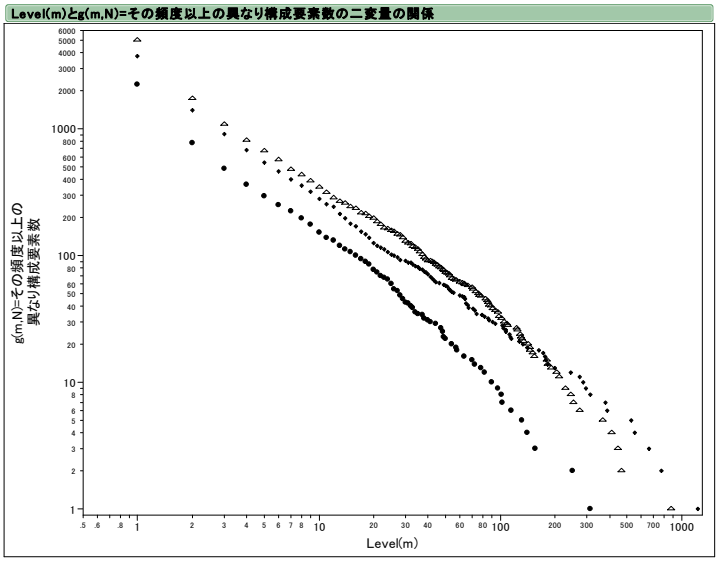

図 7　$\log m$ と $\log g(m,N)$ の関係

記号の対応：質問 2 の2調査サイト（●＝T 社、△＝D 社），質問 1 の「大切なもの」（◇＝T 社）

例３：

文芸作品を例として，上の法則を調べてみよう．夏目漱石が執筆した「坊っちゃん」「道草」「こころ」「草枕」の 4 作品について，上に示した各値を WordMiner を用いて求め，その結果をエクスポートして必要な統計値を一覧とする．それに基づいて，それぞれ図を作成する．ここでは，横軸に $\log m$ を，縦軸に $g(m,N)$ を，また横軸に順位 z の対数 $\log z$ を縦軸に頻度の対数 $\log f_z(z,N)$ を当てた図を描いてみる．

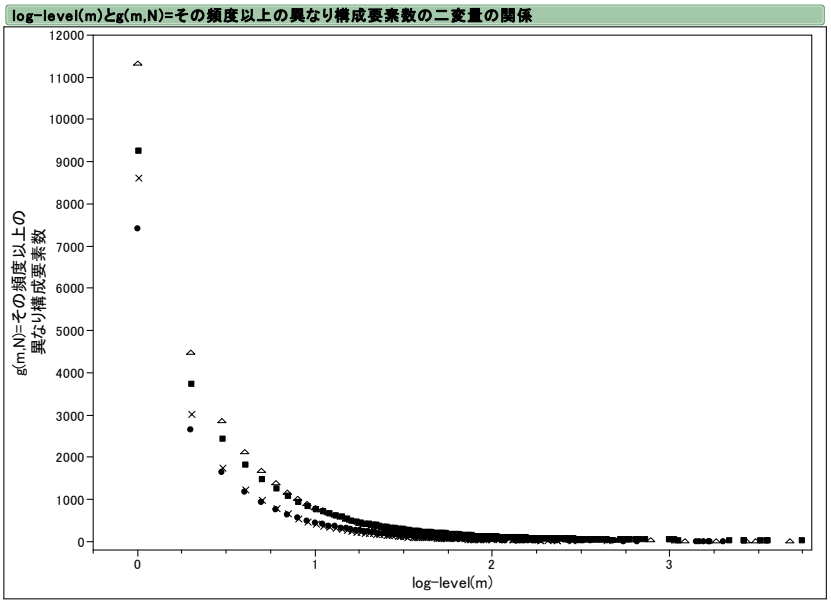

図 8　$\log m$ と $g(m,N)$ の関係

記号の対応：●＝坊っちゃん、△＝道草、◆＝こころ，×＝草枕

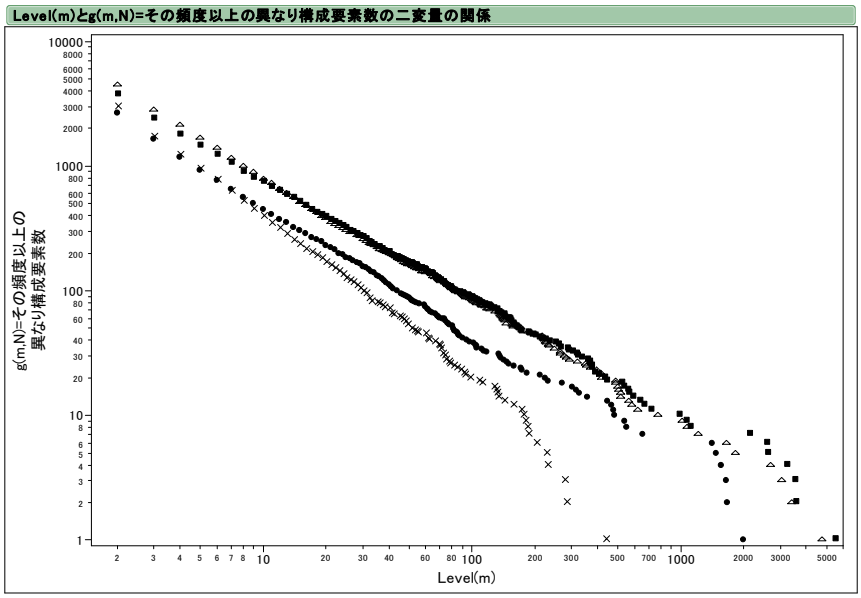

図 9　$\log m$ と $\log g(m,N)$ の関係

記号の対応：●＝坊っちゃん、△＝道草、◆＝こころ，×＝草枕

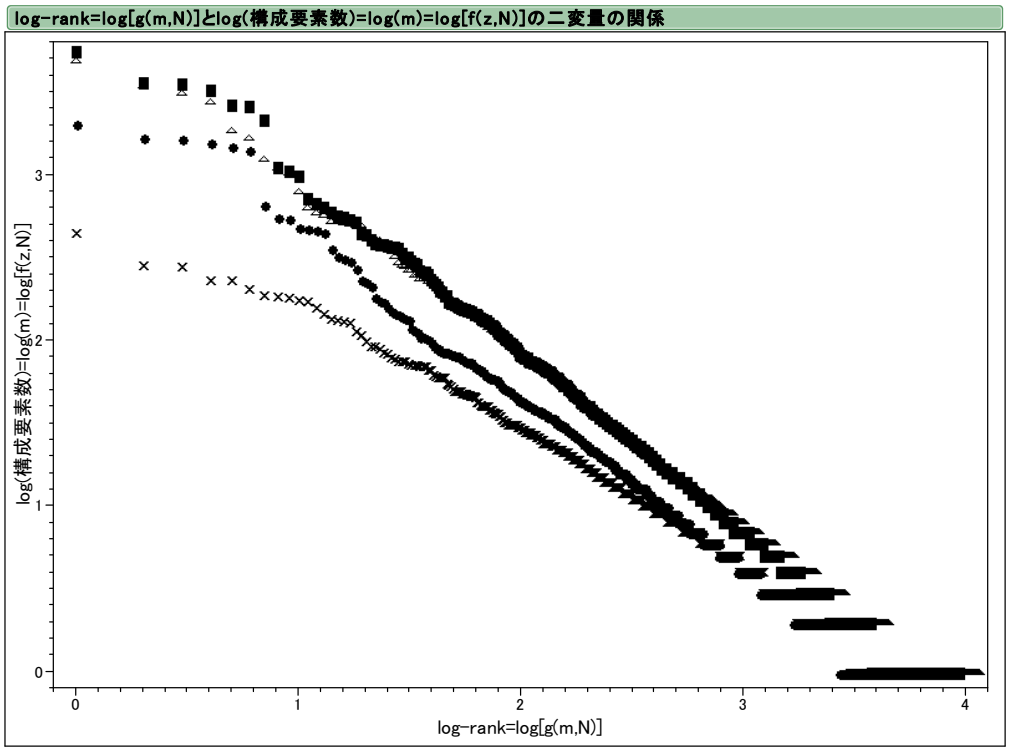

図 10　順位 z の対数 $\log z$ と頻度の対数 $\log f_z(z,N)$ の関係

記号の対応：●＝坊っちゃん、△＝道草、◆＝こころ，×＝草枕

これらの図には，以下の特徴がある．

① 4 作品の単語の出現頻度の傾向には若干の違いがみられるようだ．
② 例 2 と同様に，構成要素の出現頻度とその順位と間には類似した関係がある．順位の増加に伴って，出現単語数（構成要素数）の頻度は指数的に低減する．
③ その特性は，両軸対数のグラフからみるように式 (7) にあるような線型的な関係である．
④ ただこの例は，例 2 に比べ，裾つまり閾値 m が大きい部分では線形性が崩れている．これが総単語数（総構成要素数）の大きさの違いによるものかどうかは分からない．
⑤ ただし，例 2 では「助詞の削除」「記号類の削除」を行ったが，ここでは「記号類の削除」のみで「助詞の削除」は行っていないという編集上の違いがある（これが関係したかどうかまではさらなる分析が必要）．
⑥ 図 9 と図 10 は実質的には同じ情報である，つまり，図 10 の縦軸，横軸を入れ替えれば，実は図 9 と同じ情報に相当する（順位の付与方法によるズレは生じるが意味は同じ）．これは式（5）の相互関係から明らかである．

上の 3 つの例で示したように，構成要素（単語相当）とその出現頻度数，つまり WordMiner でいう閾値に対する構成要素数，異なり構成要素数の間には，共通した規則性があり，これを知っておくことは構成要素分布の特性を知る意味で必要ではある．しかしこのことが，文芸作品の相互比較（例えば漱石 4 作品の比較）や，自由回答質問の特徴の何を測っているのかの吟味は別の分析あるいは二次情報が必要であろう．

この他の統計的指標（計量指標）として，ユールの係数（Yule, 1944）やシンプソンの係数（Simpson, 1949），語彙潤沢度（*R.V.*: richness of vocabulary）などが知られている．これらを上に用いた記号で表すと以下のようになる．

＜ユールの係数＞

$$K = 10^4 \frac{\sum_m m^2 V(m,N) - N}{N^2} \qquad (8)$$

＜シンプソンの係数＞

$$D = \sum_m V(m,N) \frac{m}{N} \bullet \frac{m-1}{N-1} \qquad (9)$$

＜語彙潤沢度（***R.V.*: richness of vocabulary**）＞

$$R.V. = \frac{V(m,N)}{N} \qquad (10)$$

これを上に挙げた例について実際に求めたところ表 4 のようになった．ユールの係数とシンプソンの係数は，実は非常に類似しており，グラフに表すときれいな直線関係にある．この 2 つの指標と語彙潤沢度の関係をどう解釈するかはあまり明かでない．

表4　4 つの指標の一覧

指標	*Richness of vocabulary*	*Yule* の係数	*Simpson* の係数
式 ＼ 対象	$R.V. = \frac{V(m,N)}{N}$	$K = 10^4 \frac{\sum_m m^2 V(m,N) - N}{N^2}$	$D = \sum_m V(m,N) \frac{m}{N} \bullet \frac{m-1}{N-1}$
坊っちゃん	0.2735	116.2895	0.01163
草枕	0.1378	132.3211	0.01323
道草	0.1489	25.22731	0.00252
こころ	0.1190	156.0895	0.01560
大切なもの	0.1820	102.8568	0.010286
T 社	0.2604	53.46348	0.005347
D 社	0.2109	46.91445	0.004692

この節で述べたような計量化あるいは指標化は，構成要素の頻度分布の比較観察に利用できる．しかし多次元データ解析とは異なり，データに内在する潜在的な特徴の抽出・探査を直接行う方法ではないが，テキスト型データの基本的な特性を客観的に知るために必要な操作であり，WordMiner が分析過程で出力する情報を用いた事前分析の一つとして行うことは必要かもしれない．

2. 有意性テストの意味と解釈

WordMiner に用意された「**有意性テスト**」にはいくつかのオプションがある．

① 頻度による有意性テスト

これには以下のような機能を含む．

- 「(構成要素) × (質的変数)」のデータ表から出発した場合：質的情報（質的変数，属性など）と，出現した構成要素群の間の関係を調べるテスト
- 「(回答) × (構成要素変数)」のデータ表で回答・サンプルのクラスター化を行った場合：クラスター化で得たクラスター変数と構成要素群の関係を調べるテスト
- 上の検定結果をサンプル・回答単位で要約した結果の表示を行うとき

② 距離による有意性テスト

距離を用いた評価の具体的な方法の説明は，今回は省略する．意味解釈の要領だけを以下

に要約しておく．

- ここで用いる距離とはカイ二乗距離（＝χ^2距離）であり，これを用いたテストを行う（ユークリッド距離ではないこと）．
- 距離を用いるので，値が小さいほど類似性が高いと考える．
- 出発行列が（質的変数）×（構成要素変数）のとき，用いた質的変数，たとえばある質問について，その質問文内の各選択肢内でのある回答・サンプルのプロフィルが，その選択肢全体のプロフィルとどれだけ乖離しているかを測る．よって，その質問と選択肢との中での回答のまとまり具合を知る指標となる（距離の値が小さければ，その選択肢内でよくまとまっていると解釈する）．
- つまり「回答文としてその選択肢内で似ている回答」の距離が小さいとして，一覧表の上位に現れる．
- 「(クラスター変数）×（構成要素変数)」のデータ表から出発の場合は，そのクラスター変数を質的変数の一つと読み替えれば上に同じことである．つまり，あるクラスター内での回答（のプロフィル）のまとまりの程度を距離として測ることができる．
- WordMiner は，これらを「**距離による有意性テスト要約：有意なサンプルの要約**」「**距離による有意性テスト：サンプル別一覧**」として表示する（下記の図 11）．

有意性テストの結果の全体は，以下の図にみるように実行した多次元データ解析タスクのパレット内で確認できる．図 11 は，「(構成要素変数）×（質的変数)」のデータ表から出発した場合の例を示している．

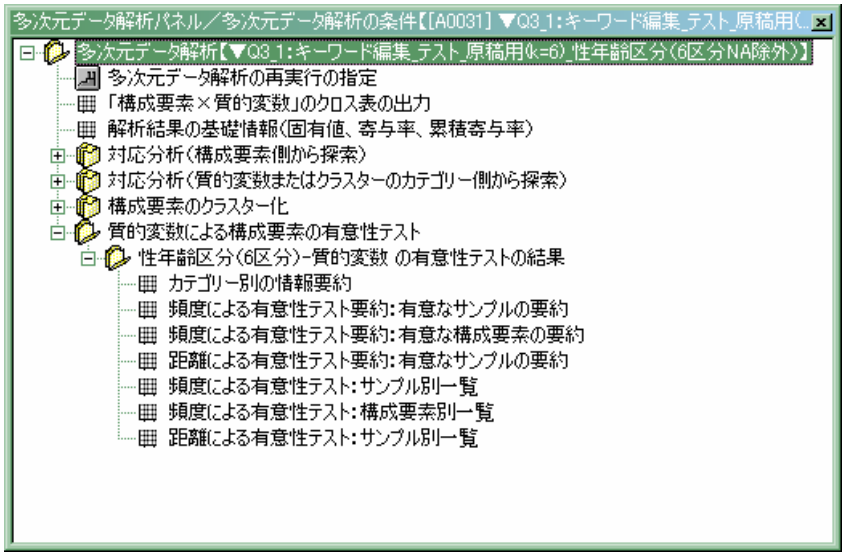

図 11 （構成要素変数）×（質的変数）の実行パレット例

2．1 頻度による有意性テスト

ここではとくに「**頻度による有意性テスト**」について，例を使ってその機能の概要を説明する．解析時の出発行列（データ表）として「(構成要素）×（質的変数)」を例として述べる．ここで出発行列が「(サンプル・回答）×（構成要素)」としてクラスター化を行った場合には，生成されたクラスター変数を，新たに得られた質的変数と読み替えれば，数理的な原理・仕組みは同じであるから，「(構成要素）×（質的変数)」についての理解が得られれば考え方は同じである．

WordMiner では，多次元データ解析を行った結果，各タスクは図 12 の「**多次元データ解析の実行内容の管理**」画面で確認する．またこの一覧内で各タスクの先頭にある「丸カラーマーク」と一覧説明で，「(構成要素）×（質的変数)」「(サンプル）×（構成要素)」のいずれであるかを識別できる（図 12 の「種類」の欄）．

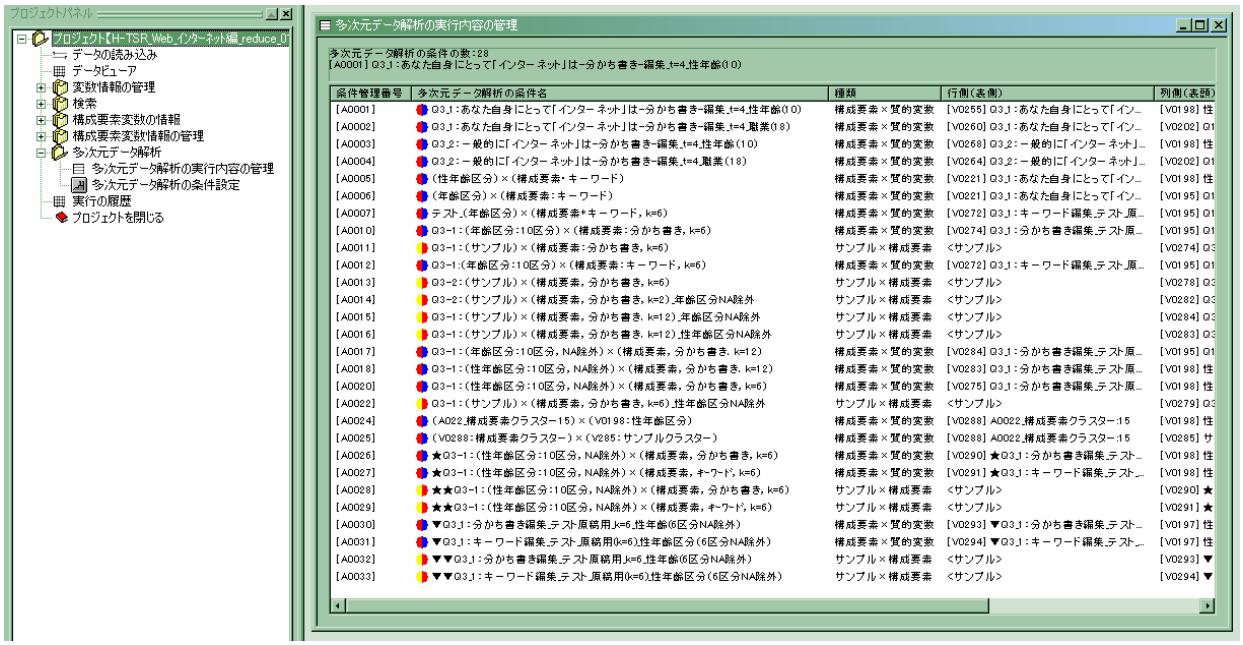

図 12 「多次元データ解析の実行内容の管理」画面の例

2．2 有意性テストで扱う情報

「**頻度による有意性テスト**」では，図 11 に見るように以下の 4 種類の情報が出力される．

I　頻度による有意性テスト：構成要素別一覧
II　頻度による有意性テスト：サンプル別一覧
III　頻度による有意性テスト：有意なサンプルの要約
IV　頻度による有意性テスト：有意な構成要素の要約

この並び順は図 11 の表示と順序が異なるが，この中で基本となる情報は **I** の「**頻度による有意性テスト：構成要素別一覧**」である．まずこの一覧で得られる情報の内容を説明する．Ⅱ～Ⅳは，Ⅰで得られた情報の要約情報である．

また以下に示す説明から，これらの各一覧情報の相互の関係をよく理解して使い分けることが肝要である．

2．3 頻度による有意性テストの仕組み

有意性テストを理解するには，初等統計学の若干の知識を必要とする．例えば以下のような事項である．

- 統計的検定法（検定統計量と検定統計値，仮説検定，有意水準など）
- 有意確率
- 標準化の操作
- 超幾何分布とその特徴（母数など）
- 超幾何分布の正規近似
- 超幾何分布，二項分布の正規近似
- 母集団と標本，非復元抽出（sampling without replacement）

こうした知識をここで記述することには紙幅の都合で無理があるので，以下では，これらの知識がある程度あることを想定して記述する．

まず以下の記号・記法を準備する.

$$\begin{cases} k: \text{総構成要素数} \\ k_{i+}: \text{ある構成要素}i\text{の出現頻度} \\ k_{+j}: \text{ある対象群・層}j\text{内の構成要素数} \\ k_{ij}: \text{ある構成要素}i\text{がある対象群・層}j\text{内に出現する度数} \end{cases}$$

ここで“ある対象群・層”とした箇所は，質的変数，例えば選択肢型質問であればその個々の選択肢(カテゴリー)を示す. クラスター変数であれば, 各クラスターを指す. ここでは表側(ひょうそく)に構成要素群を置き，表頭(ひょうとう)に対象とする群・層を置いてある．例えば属性として「性年齢区分」を考えると，個々の性年齢区分の選択肢に相当すると考える．これらの関係は次の図 13 と併せて考えると理解が容易である．

「**頻度による有意性テスト：構成要素別一覧**」から得られる一覧情報は，ここで各選択肢別の表 5 の形式の情報を表示したものである．例えば，ここの例であると，性年齢区分の 12 カテゴリーについて表 2 のシートが 12 枚出力表示される（後の表 6 がその一例）.

表 5　構成要素他の考え方

		対象群・層（例：ある質的変数とその選択肢）						
		1	2	…	j	…	n	和（構成要素ごとの和）
異なり構成要素の並び	1	…	…	…	k_{1j}	…	…	k_{1+}
	2	…	…	…		…	…	…
	…	…	…	…		…	…	…
	i	k_{i1}	…	…	k_{ij}		k_{in}	$k_{i+}\left(=\sum_{j=1}^{n} k_{ij}\right)$
	…	…	…	…		…	…	…
	V	…	…	…	k_{Vj}	…	…	k_{V+}
	和（選択肢別の和）	…	…	…	$k_{+j}\left(=\sum_{i=1}^{V} k_{ij}\right)$	…	…	$k\left(=\sum_{i=1}^{V} k_{i+}=\sum_{i=1}^{V}\sum_{j=1}^{n} k_{ij}\right)$（総構成要素数）

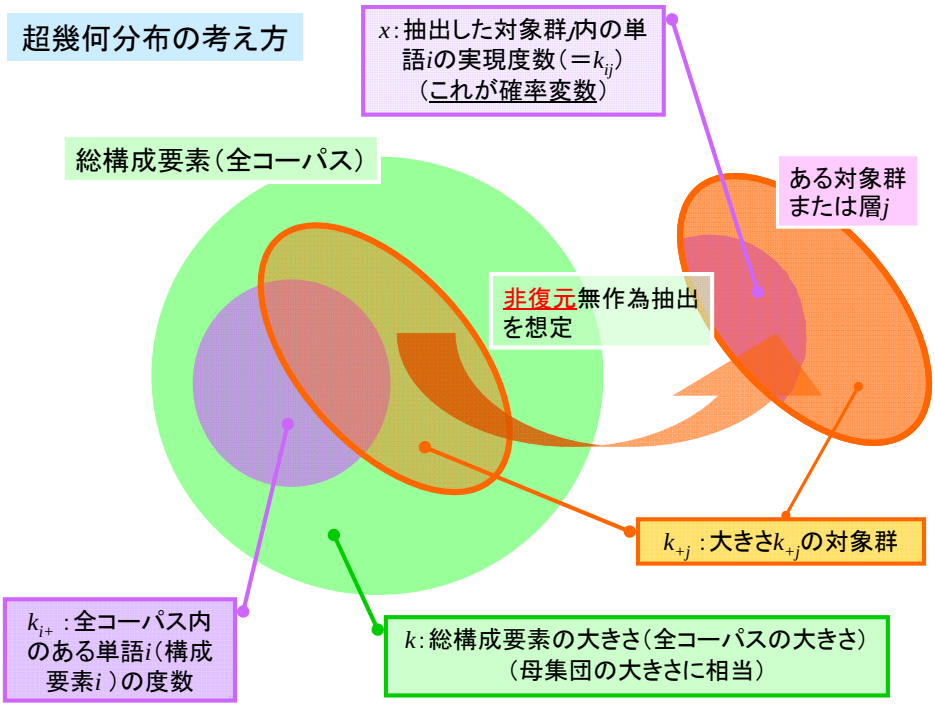

図 13　超幾何分布における構成要素の関係

以上の準備のもとに．以下のように考える．

① まず，対象としたデータセット内の総構成要素数（つまり分析対象課題のコーパス）を（有限）母集団のように考え，その大きさを k とし，ここに含まれる構成要素 i の頻度を k_{i+} とする．

② このコーパスから非復元抽出で，ある対象群・層 j に注目し，この j 内に含まれる全構成要素数 k_{+j} を抽出したとする．

③ このとき，ある構成要素 i がある対象群・層 j 内に出現した頻度数 k_{ij} を確率変数と考え，またこの確率変数が超幾何分布に従うものと考える．

④ つまり，この k_{ij} の示す確率分布を以下の超幾何分布で表わすものとする．

$$P\left(x=k_{ij}\right)=\frac{\binom{k_{i+}}{x}\binom{k-k_{i+}}{k_{+j}-x}}{\binom{k}{k_{+j}}} \quad \begin{pmatrix} x=0,1,2,\cdots,t \\ t=\min\left\{k_{i+},k_{+j}\right\} \end{pmatrix} \tag{11}$$

⑤ この超幾何分布の期待値（理論上の平均値，母平均）と分散（母分散）は以下のようになる．

$$m(i,j)=\frac{k_{i+}k_{+j}}{k} \quad \begin{pmatrix} x=k_{ij}\text{に対する期待度数} \\ \text{ある対象群・層}j\text{内の総構成要素数}k_{\cdot j}\text{中} \\ \text{に占める単語}i\text{の期待される出現度数} \end{pmatrix}$$

$$s^2(i,j)=\frac{k_{i+}k_{+j}}{k}\left(1-\frac{k_{i+}}{k}\right)\left(\frac{k-k_{+j}}{k-1}\right) \quad \left(x=k_{ij}\text{の分散}\right) \tag{12}$$

$$s(i,j)=\sqrt{\frac{k_{i+}k_{+j}}{k}\left(1-\frac{k_{i+}}{k}\right)\left(\frac{k-k_{+j}}{k-1}\right)} \quad \left(x=k_{ij}\text{の標準偏差}\right)$$

⑥ この確率変数 k_{ij}（実現度数）の標準化変数を正規近似するとした次の検定統計量を作る．

$$T(i,j)=\frac{k_{ij}-m(i,j)}{s(i,j)} \simeq N\left(0,1^2\right) \quad \begin{pmatrix}\text{近似的に標準化変数が}\\ \text{標準正規分布}N\left(0,1^2\right)\text{に}\\ \text{従うとして検定する}\end{pmatrix} \qquad (13)$$

⑦ 実際に得られた個々の k_{ij} について，上の式により検定統計量 $T(i,j)$ の実現値である検定統計値（検定値） $t(i,j)$ を求めてテストを行う．

⑧ たとえば，このときの有意確率や有意水準を与えたときの棄却域 R は以下のように考える．

(i) $T(i,j)$の実測値，つまり得られた検定値を$t(i,j)$としたとき

有意確率$\left(p\text{値}\right)$は以下となる．

$$P\left(\left|T(i,j)\right|\geq t(i,j)\right)=p$$

(ii) 有意水準$\alpha=0.05$に対する（両側）検定の棄却域Rは以下となる．

$$P\left(\left|T(i,j)\right|\geq 1.96\right)=0.05 \;\Leftrightarrow\; R=\left\{T(i,j)\geq 1.96 \quad \text{または} \quad T(i,j)\leq -1.96\right\}$$

注1） WordMiner ではこの検定に用いる判定基準として，例えば有意水準 $\alpha=0.05$ に対する数値 1.96 を 100 倍値で与えるようになっている（つまり 196 と指定する）．もちろん他の値をフィルターとして与えることもできる．また出力したい表示個数で与えることも可である．しかし実用上は「有意確率」を参考に解釈を行うことがよい．

注2） 超幾何分布は，k，k_{i+} が非常に大きければ二項分布で近似できる．

ここでの考え方は，以下のようになる．

① 総構成要素数（全コーパス）に占めるある構成要素 i の頻度が占める比率（出現比率），つまり $p=\frac{k_{i+}}{k}$ が，分かったとする．（注：検定に現れる有意確率 p 値とは別の意味）

② このとき，ここから抽出したある層 j 内の同じ構成要素 i の出現頻度，つまり k_{ij} が超幾何分布に従って分布すると仮定したとき，どの程度の（推定）頻度となるか（上の期待値）を考える．

③ その推定した期待値（予想される出現頻度）をさらに正規近似した検定統計量を考え，その検定統計値（検定値）でテストする．

④ つまりそれがゼロに近ければ有意でないし（総構成要素数つまり全コーパス k 内に占める k_{i+} の傾向，出現比率 p の傾向に似ている），かなり大きな値となるならば全コーパス内にみられるある構成要素 i の傾向とは異なることになる．

⑤ よって，検定統計値とそれに対する有意確率（p 値）を出力し観察することで構成要素 i のその層 j 内での特徴的な傾向を測ること，つまりその層内の特徴的な構成要素を客観的に知ることができる．

⑥ この他，構成要素の出現比率，つまり $k_{i+}/k, k_{ij}/k_{+j}$ などの分布を観察する（後述の例参照）．

２．４　数値例による説明

簡単な数値例を挙げておく．上に示した各記号を対応させておくので，両者を比べて観察するとよい．

例題：

ある Web 調査で用いた自由回答質問と，質的変数として「性年齢区分」を用いた例を示す．

＜調査内容＞

ここでは，調査方式として Web 調査を用いた．分析対象とした回答者数は「812（サンプル)」，扱う異なり構成要素数は「148 語（=*V*)」，全構成要素数は「3152 語（=*k*)」であった．従って，異なり構成要素率は，148/3152≒4.7（%）となる．

つまりデータセットとしての規模（大きさ）はさほど大きくはない．分析上はこのことを十分に念頭において以下を解釈することが肝要である．

＜自由回答質問＞

「あなたご自身にとって「インターネット」は、どのようなことがらに活用できると思いますか。どんなことでも結構ですので、以下になるべく具体的にご記入ください。」

＜性年齢区分（質的変数）の 12 の選択肢：*n*=12＞

これは以下の 12 選択肢とする（NA は除外した)．

1　女性/20 才未満
2　女性/20 才～29 才
3　女性/30 才～39 才
4　女性/40 才～49 才
5　女性/50 才～59 才
6　女性/60 才以上
7　男性/20 才未満
8　男性/20 才～29 才
9　男性/30 才～39 才
10　男性/40 才～49 才
11　男性/50 才～59 才
12　男性/60 才以上

サンプル数，用いた構成要素数からみて，この選択肢数はやや多めであることを注意しておこう．このデータからいま，選択肢 9 の「男性/30 才～39 才」を選ぶ．このサンプル数も 124 (名）とさほど大きくはない．ここで上の表に合わせると選択肢総数 *n*=12 (カテゴリー)，*j*=9 と対応する．以下の議論は他の選択肢に対しても，同じように適用できる．また WordMiner では，これを各年齢区分別にそれぞれ一覧としてシートに出力する．

ここでは，選択肢（つまり層）*j*＝「男性/30 才～39 才」としたときに得られた情報と，総構成要素数（全コーパス）との関係を WordMiner の結果表「**頻度による有意性テスト：構成要素別一覧**」から取り出すと表 6 のようになる．WordMiner では，エクスポート機能により全情報をテキスト・ファイル（csv ファイル）として出力する．まず WordMiner が表示・出力するこの一覧の見方に慣れる必要がある．

この表 6 に現れる構成要素について，ここで取り上げた層＝「男性/30 才～39 才」では，どのように説明力を持つものかを示す指標が検定値である．

例えばここで数値計算例として，構成要素として「情報収集-情報集め」「コミュニケーション」「旅-旅行」の 3 個を選んでみよう．検定値の大きさとその符合から，始めの 2 個はこの層（男性/30 才～39 才）で有効に働く，意味がある構成要素であり，最後の一つは，この層

内ではあまり説明力がなさそうな構成要素ということを示している．
以下でこの 3 つの構成要素について，実際に上の式に従って検定値を求めよう．

数値例 1：
まず共通情報として以下がある．

$$\begin{cases} k = 3152\left(\text{総構成要素数}\right) \\ k_{+j} = 385\left(\text{選択肢}j = \text{「男性/30才～39才」内の構成要素数}\right) \end{cases}$$

ここで構成要素「情報収集−情報集め」について検定統計値を算出する．
$k_{i+} = k_{1+} = 135$，$k_{ij} = k_{1j} = 31$ に注意して以下のように求める．

$$m(i,j) = \frac{k_{i+}k_{+j}}{k} \Rightarrow m(i,j) = \frac{135\times 385}{3152} \fallingdotseq 16.48953$$

$$s^2(i,j) = \frac{k_{i+}k_{+j}}{k}\left(1-\frac{k_{i+}}{k}\right)\left(\frac{k-k_{+j}}{k-1}\right) \Rightarrow \frac{135\times 385}{3152}\left(1-\frac{135}{3152}\right)\left(\frac{3152-385}{3152-1}\right) \fallingdotseq 13.85984$$

$$s(i,j) = \sqrt{\frac{k_{i+}k_{+j}}{k}\left(1-\frac{k_{i+}}{k}\right)\left(\frac{k-k_{+j}}{k-1}\right)} \Rightarrow s(i,j) = \sqrt{\frac{135\times 385}{3152}\left(1-\frac{135}{3152}\right)\left(\frac{3152-385}{3152-1}\right)} = 3.72288$$

$$T(i,j) = \frac{k_{ij} - m(i,j)}{s(i,j)} \Rightarrow t(i,j) = \frac{31-16.48953}{3.72288} \fallingdotseq 3.898$$

数値例 2：
同じようにして，構成要素「コミュニケーション」について求めると次を得る．

$$m(i,j) \fallingdotseq 3.66434,\quad s^2(i,j) \fallingdotseq 3.18715,\quad s(i,j) \fallingdotseq 1.78526$$

$$t(i,j) = \frac{9-3.66434}{1.78526} \fallingdotseq 2.9887$$

数値例 3：
同じようにして，構成要素「旅-旅行」についても求めた．

以上で求めた数値を，WordMiner で求めた値と表 7 に一覧とした．ここで，WordMiner の出力値と手計算による値の違いは WordMiner が上に示した数式とは異なるある数値計算アルゴリズム（近似計算）を用いていることから生じるものである．
誤差が大きいと感じるかもしれないが，この検定値の利用目的は序列（検定結果の並び）あるいは検定統計値の大きさの順に関心があるのでとりあえずこれで問題はない．ただし，構成要素数の情報のそれぞれの出現頻度数が少ない場合や，均衡がよくないときには，やはり注意する必要がある（後述）．
この例の場合の有意性テストの結果解釈は以下のようになる．

① ここで観察した層，つまり性年齢区分が「男性/30 才～39 才」の層を特徴付ける構成要素として，検定値の上位をみると「情報収集ー情報集め」「「仕事」「情報検索」「インターネットショッピング-オンラインショッピング」「コミュニケーション」「新聞」「インターネットバンキング」「関係」「情報源」「通信手段」などがある．
② 一方，この性年齢区分「男性/30 才～39 才」にとっては下位にある「友達-友人」「購入」「旅-旅行」などはあまり説明力があるとはいえない．
③ 併せて構成要素の出現比率，つまり $k_{i+}/k, k_{ij}/k_{+j}$ を観察する．この操作は重要である（検

定値の観察より先に行うべきである)．この例では例えば上位 1 位の構成要素「情報収集－情報集め」は欄⑤の比率 $\frac{k_{1j}}{k_{+j}}=\frac{31}{385}=8.05(\%)$ と欄⑥の比率 $\frac{k_{1+}}{k}=\frac{135}{3152}=4.28(\%)$ を比べるとこの構成要素がこの性年齢区分「男性/30 才～39 才」で特徴的であることが分かる．他の構成要素についてもこの比率の比較に意味がある．

表 6　選択肢「男性/30 才～39 才」に対する頻度の有意性テスト結果

		①	②	③	④	⑤	⑥	⑦	⑧
		検定順位	キーワード編集	検定値	有意確率	カテゴリー内（構成要素数構成比）	構成要素数（構成比）	ｶﾃｺﾞﾘｰ内構成要素数	構成要素数
異なり構成要素 i の情報に対応	i	一部を表示	抽出した構成要素	$t(i,j)$	p 値	⑦÷385×100	⑧÷3152×100	k_{ij}	k_{i+}
	1	上位 1	情報収集-情報集め	**3.47**	0	8.05	4.28	**31**	**135**
	2	上位 2	仕事	2.84	0	4.68	2.32	18	73
	3	上位 3	情報検索	2.8	0	1.56	0.41	6	13
	4	上位 4	インターネットショッピング-オンラインショッピング	2.68	0	1.82	0.57	7	18
	5	上位 5	コミュニケーション	**2.43**	0.01	2.34	0.95	**9**	**30**
	6	上位 6	新聞	2.02	0.02	1.56	0.6	6	19
	7	上位 7	インターネットバンキング	1.92	0.03	0.78	0.19	3	6
	8	上位 8	関係	1.92	0.03	0.78	0.19	3	6
	9	上位 9	情報源	1.53	0.06	0.78	0.25	3	8
	10	上位 10	通信手段	1.53	0.06	0.78	0.25	3	8
	…	…	＜途中は省略＞	…	…	…	…	…	…
	11	下位 3	友達-友人	-1.66	0.05	1.3	2.63	5	83
	12	下位 2	購入	-2.01	0.02	0	0.92	0	29
	13	下位 1	旅-旅行	**-2.58**	0	0.26	1.81	**1**	**57**
						(100)	(100)	**385** ($=k_{+j}$)	**3152** ($=k$)

表7　検定値の算出例

		①	②	③		④	⑦	⑧
		検定順位	キーワード編集	検定値（数値計算）	手計算による検定値	有意確率（前に同じ）	ｶﾃｺﾞﾘｰ内構成要素数	構成要素数
異なり構成要素 i	i	一部を表示	抽出した構成要素	$t(i,j)$	$t(i,j)$	p 値	k_{ij}	k_{i+}
	1	上位 1	情報収集-情報集め	**3.47**	**3.898**	0	**31**	**135**
	5	上位 5	コミュニケーション	**2.43**	**2.989**	0.01	**9**	**30**
	13	下位 1	旅-旅行	**-2.58**	**-2.066**	0	**1**	**57**
							385 ($=k_{+j}$)	**3152** ($=k$)

なおここでは，出力表のごく一部を眺めたが，実際には「出力個数」の制御や判定基準とする検定値などを変えて入力し結果を総合的に観察することが望ましい．出力個数のボリュームを気にしないなら，なるべく多くを一覧として観察するとよい．

このとき，同じ値の検定値が並ぶことがあり（頻発する），また検定の仕組みから考えて構成要素数（k_{i+}，k_{+j}，k_{ij} など）が極端に少ない構成要素の解釈には注意する．

2．5「頻度による有意性テスト:有意な構成要素の要約」の内容

上でみた「**頻度による有意性テスト：構成要素別一覧**」で得られる各シートの情報から，有意性テストで得た各層の構成要素を順位に合わせて一覧要約した表が「**頻度による有意性テスト：有意な構成要素の要約**」である．

上に挙げた例の場合，（順位でソートした構成要素）×（質的変数の選択肢，つまり性年齢区分）の形式の要約表を出力する．選択肢別の表で，検定値から構成要素の個々の意味・説明力を観察した後，こちらの一覧情報と比較観察すると，構成要素の意味解釈がより容易になる．

ここで用いた例について，表の一部を示すと図 14 のようになる．この例では，検定値の基準を「100」つまり「1.00」として出力して得られた表の一部である．もちろん他の基準値や出力したい個数（抽出数）を指定することができる．

	女性/1_20才未満 サンプル数：23 異なり構成要素数：51	女性/2_20才～29才 サンプル数：107 異なり構成要素数：112	女性/3_30才～39才 サンプル数：152 異なり構成要素数：128	女性/4_40才～49才 サンプル数：89 異なり構成要素数：112	女性/5_50才～59才 サンプル数：39 異なり構成要素数：84	女性/6_60才以上 サンプル数：17 異なり構成要素数：36	男性/1_20才未満 サンプル数：16 異なり構成要素数：26	男性/2_20才～29才 サンプル数：75 異なり構成要素数：88	男性/3_30才～39才 サンプル数：114 異なり構成要素数：107	男性/4_40才～49才 サンプル数：99 異なり構成要素数：98	男性/5_50才～59才 サンプル数：46 異なり構成要素数：81	男性/6_60才以上 サンプル数：35 異なり構成要素数：60
上位 1	学校	交流	子育て	参考	病院	旅-旅行	事-こと	人	情報収集-情報集め	仕事上	各種	航空券
上位 2	サイト	アーティスト	下調べ	利用	電話	病気	勉強	簡単	仕事	情報収集-情報集め	手紙	会社
上位 3	レポート	チェック	インターネットオークション	友達-友人	電子メール-メール	知人-しりあい	ネット上	サイト	情報検索	仕事	情報入手	参加
上位 4	懸賞	場所	病気	ため-為	ヒント	時刻表	宿題	交流	インターネットショッピング	調査	仕事上	旅行先
上位 5	学習	家	旅-旅行	勉強	天気予報	買い物	調べ物	買い物	コミュニケーション	航空券	百科事典	予約
上位 6	資料収集	資料収集	気	銀行	電車	交換	連絡手段	掲示板	新聞	自宅	電話	ホテル
上位 7	色々-いろいろ	場合	料理	子供	意見	行き先	ゲーム	学習	インターネットバンキング	収集	ヒント	取引
上位 8	電車	事-こと	子供	趣味	時刻表	交通機関	ダウンロード	気軽	関係	予約	調査	交通機関
上位 9	時刻表	手軽	電車	手段	私	使用	音楽	連絡手段	情報源	生活	連絡手段	書籍
上位 10	趣味	売買	時間	学習	知識	日常生活	自分	チャット	通信手段	インターネットバンキング	発信	情報交換
上位 11	興味	お店	ホテル	日常生活	百科事典	病院	学校	レポート	趣味	価格	知識	情報検索
上位 12	自分	人	レシピ	余暇-レジャー	色々-いろいろ	時刻	辞書代わり	商品	音楽	交換	メール交換	商品
上位 13	ホームページ	辞書代わり	交通手段	興味	料理		チャット	購入	人	通信販売-通販	情報発信	購入
上位 14		ホームページ	写真	レストラン	手紙		活用	コミュニケーション	ため-為			インターネット-ネット
上位 15		共通	手軽	時刻	事柄				情報発信			地図
上位 16		専門的	旅行情報	便利	辞書				調査			チケット
上位 17		地図	確認	収集								確認
上位 18		インターネット-ネット	友達-友人	情報								
上位 19		ニュース	言葉									
上位 20		レストラン	場所									
上位 21		企業										
上位 22		書籍										

図 14　頻度による有意性テスト:有意な構成要素の要約の例

2．6　頻度による有意性テスト:サンプル別一覧

「**頻度による有意性テスト：構成要素別一覧**」で得られる情報をサンプル単位で観察するための情報が「**頻度による有意性テスト：サンプル別一覧**」に表示される．

ここには，「**頻度による有意性テスト：構成要素別一覧**」の表示に合わせて，個別のシートに情報が出力される．例えばここでみた例では，質的変数である性年齢区分別の選択肢の 12（カテゴリー）のシートが出力される．上で有意性テストの仕組みの説明に用いた性年齢区分「男性/30 才～39 才」に対応するシートを示すと図 15 となる．ここでは，抽出数を 10 としたので，検定値の大きさに従い上位の 10 サンプルを出力した．

図 15　頻度による有意性テスト:サンプル別一覧，性年齢区分が「男性/30 才～39 才」の出力例

この一覧に表示されている「検定値」は，上に表 6，表 7 に示したそれとは異なる情報である．これについて説明する．

表 8　図 15 の抜粋（書き替えた表）

サンプル	検定値	Q3_1：キーワード編集
1	3. 47	情報収集-情報集め
2	3. 47	情報収集-情報集め
3	3. 47	情報収集-情報集め
4	3. 47	情報収集-情報集め
5	3. 47	情報収集-情報集め
6	3. 47	情報収集-情報集め
7	2. 91	情報収集-情報集め コミュニケーション 仕事
8	2. 80	情報検索
9	2. 45	情報収集-情報集め 趣味
10	2. 16	情報検索 通信手段

いま図 15 の情報を表 8 のように整理した．サンプルラベルとしてあった「SEQ」を除き得られた 10 サンプルを表示してある．この情報は，この選択肢（男性/30 才～39 才）で有意となった構成要素を用いて以下のように算出したサンプル別の「検定値」の大きさで並べ替えたものである．ここでいう「検定値」とは，前にみた構成要素別の検定値の**調整済み平均値**(注)となっている．つまり，あるサンプルが回答した構成要素の個々に与えられた検定値を，その個数で調整し平均した値である．

例えば表の先頭にあるサンプル 1 を見ると検定値が「3.47」とある．このサンプルが示した回答が「情報収集－情報集め」の一語（1 構成要素）だけであるから，前に表 6 で得た検定値 3.47 が当てられる．

さらに別の例として，サンプル 7 を見よう．このサンプルは「情報収集-情報集め」「コミュニケーション」「仕事」と 3 語を回答している．表 6 でこれら 3 語に対する検定値をそれぞれ検索すると，

「情報収集-情報集め」　3.47
「コミュニケーション」　2.43
「仕事」　2.84

である．よってこのサンプル 7 に対する検定値（サンプル検定値）は，3 個の検定値の単純な平均値は，

（3.47+2.43+2.84）÷3≒2.91

となる．表 8（かつ図 15）の該当サンプルの検定値がこれに相当する．以下，同じように各サンプルに対する検定値を求める．こうして表 8（図 15）が得られる．

（注）調整済み検定値について

現実には，検定値が非常に小さいとき，あるいは逆に非常に大きい値となるような場合，つまり出現単語の頻度が極端に偏った大きさを示すようなことがある．しかもこれが実用上はかなり頻繁に現れる．このような場合に出現頻度による単純な算術平均を用いるとこの統計値の性質上偏った結果となる．そこで WordMiner では，ある調整を行った上での**調整済み平均値**を用いている．これは一種の経験則であって，以下のように検定値の大きさでフィルタリングを行って平均値を調整した**調整済み検定値**として用いる．

① 個々の構成要素の検定値 $t(i,j)$ の大きさを調べ，ある範囲から外れた場合それを除外した上で平均値を求める．例えば具体的には，$t(i,j)$ の絶対値について $1.0 \leqq |t(i,j)| \leqq 99.9$ の条件でフィルタリングしている．つまり，検定値が小さくてさほど有意でない構成要素と，逆に検定値が極端に大きくはずれ値のように振る舞う構成要素を除外する．これは過去の多数の分析から得た一つの経験則である．

② この調整済み平均値の算出時，分母とする出現頻度は除外した構成要素数も含めた，つまり全体の出現頻度数を変えずにそのまま用いる．

以上からここに示す一覧情報は，各サンプルが用いた構成要素を平均的にみたときの有意の程度を示す指標である．そしてある回答の中で特徴的な構成要素が多く使われているときに，この値が大きくなる．

2．7 「頻度による有意性テスト：有意なサンプルの要約」の内容

これは上でみた「**頻度による有意性テスト：サンプル別一覧**」で得られる情報を一覧表示したものである．図 16 がその出力情報の一部である．この表は表側がサンプル，表頭に指定した個数だけ表示した各選択肢，つまりここでは性年齢区分別の代表的なサンプルが並ぶ．

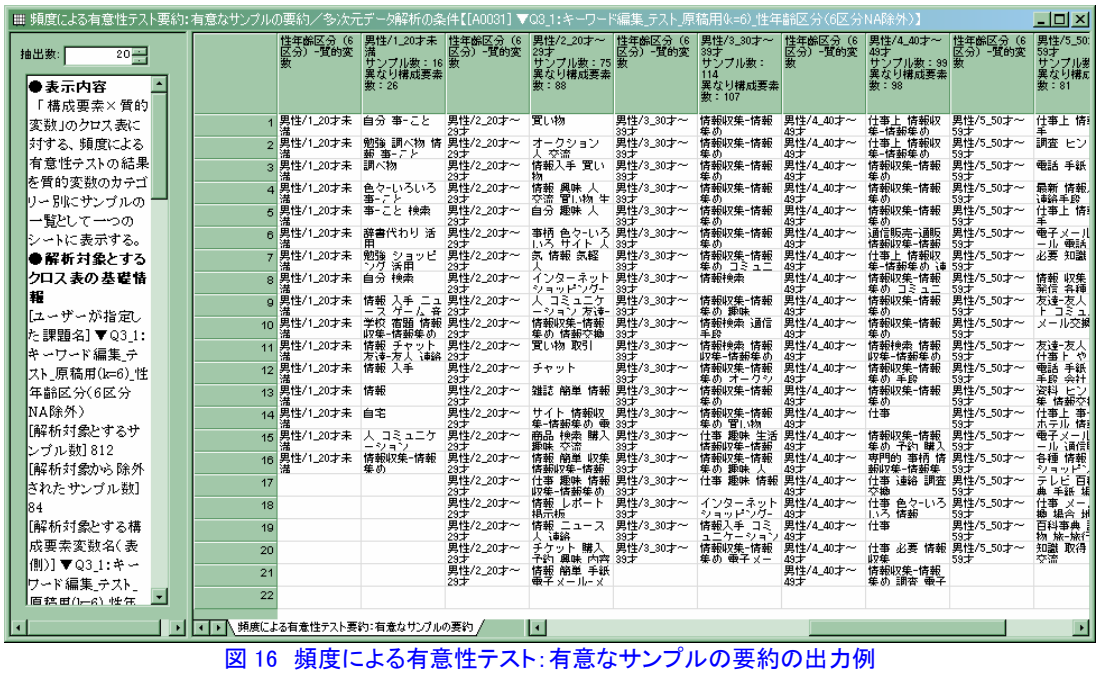

図 16　頻度による有意性テスト：有意なサンプルの要約の出力例

2．8　頻度による有意性テストの利用上の一般的な留意事項

頻度による有意性テストの結果の解釈に際して，その誘導の仕組みから，以下のような事項に留意することが必要である．

① この種の分析では扱うデータ数，つまり構成要素数，異なり構成要素数がかなり大きいサイズを想定していることがある．例えば上に示した例のサンプル数（回答者数）と構成要素数（実際には分析に用いた異なり構成要素数）の出現頻度数のいずれも，さほど大きくはない．このことから正規近似の程度に揺らぎが生じる．
② 換言すれば，分析対象とする構成要素数，異なり構成要素数はなるべく大きいことが望ましい．また層の数も不必要に大きくしない方がよい．質的変数であれば選択肢数をあまり多くしない，クラスター数であればこれもサンプル数・回答数の大きさを勘案し，それらがあまり大きくないときには細かく分類しない，といった配慮が必要である．
③ 分析対象で扱う総構成要素数，異なり構成要素数と出現頻度を観察し，どの程度のボリュームの情報を扱っているかに常に配慮すること．
④ 上に示した正規近似は，構成要素数の大きさが小さいときにはあまり近似がよくないことが知られている．検定値を大きさの順に並べて出力する理由の一つは，検定値の大きさの順に観察することに意味があるためだが，そのとき構成要素数と層内の構成要素数（記号でいうと k_{i+}，k_{ij}）の頻度の大きさと均衡に注意すること，あまり少ない値のときの検定値は注意して読みとる．
⑤ （前述のように）「**頻度による有意性テスト：構成要素別一覧**」から得られる情報のうち，実際の検定値の算出には数値計算アルゴリズムを用いている（出現頻度の大きさに合わせて何通りかの近似式を使い分けている）．よって前述の数式で得られる値との間に，正規近似の揺らぎの他に，数値計算誤差が生じることがある．

「頻度による有意性テスト」では，始めに列記した次の 4 種の情報の相互の関係を知り目的の応じて使い分けることが必要である．このテキスト資料を通じて，基本となる「**頻度による有意性テスト：構成要素別一覧**」の内容と他の出力要約情報を有効に利用していただきたい．

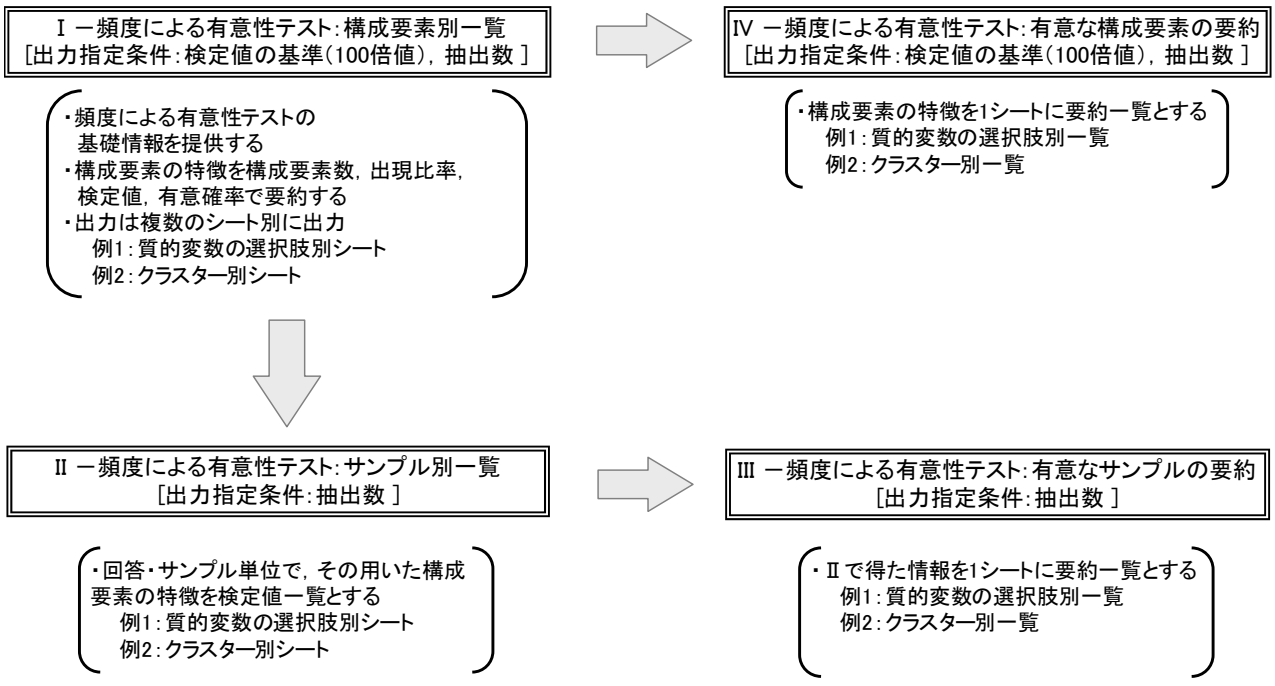

図 17　構成要素の頻度のよる有意性テストの出力結果の関係