

JMRAマーケティング・リサーチ講座

質的データのマイニング のための対応分析法

スライド資料[その1]

大隅 昇

ohsumi@ss.ij4u.or.jp

<http://wordminer.org/>

Copyright by Noboru Ohsumi

まえおき

- 本セミナー用に独自に作成の「テキスト」と「スライド資料」を用いる.
- スライド資料にトークに必要な基本事項は書き入れた.
- スライドの図表, 文字共に多いが, 聴いていただき, あとで利用できるよう配慮した(つもり).
- テキスト(該当箇所)を引用して「ことば」「図表」で説明することもある.
- 対応分析法はもはや古典的手法である. しかしこれを巡る多種多様な研究がある(発展中でもある).
- ここでは, きわめて基本的な部分と, その特性を利用したクラスター化法に話題を絞って話す.

(つづき)

- 若干の“数式”が登場する. 基礎統計学, 線形代数の入門的な知識が必要かもしれない(イメージで把握).
- 数式の細かい説明はしない. “符丁の意味”を知るために用いる(それで十分, 仕組みを理解).
- なるべく, “トイ・データ, 実際データを用いた数値例”で説明するよう努める.
- すでに周知のこと, 習得知識もあるだろうが, 一方, こういう“見方もある”と復習として聞いていただく.
- なるべく平易に“繰り返し”話す. 多少, くだくなる.

◎疑問は残さないよう, いつでも質問を!!!

◎基本的な“用語”や, 特有の“方言”に慣れていただく.

配付資料の構成

- 「テキスト」(コピー資料)として
- [第Ⅰ部]～[第Ⅲ部]を用意した(約250ページ)
 - [第Ⅰ部] 対応分析法とは(概要)
 - [第Ⅱ部] 対応分析法の基本的な考え方(基本数理)
 - [第Ⅲ部] 対応分析法とクラスター化法
- 「スライド資料」として
- いくつかに分ける:[その1], [その2], ..., [その5](予定)
- トークの進捗に合わせて内容調整し, 配布する.
- キーワード, コラム的なトピックスを先頭「★」印ページ, また知っておきたい用語などを赤枠線囲みで記す.

スライド資料[その1]の内容

- 対応分析法(CA, AFC)の誕生, 歴史的経緯.
- 基本的な考え方, 概念など(抜粋). これが重要.
- 類似手法, とくに数量化法III類(パターン分類)との関係.
- 提唱者たち(ベンゼクリ氏, 林知己夫氏)の考え方.
(注: 以下, 敬称略す)
- “データ解析”, “データの科学”の考え方.

対応分析法(CA: correspondence analysis)

正確には, AFC(Analyse Factorielle des Correspondances)あるいはAC(Analyse des Correspondances)

(つづき)

- データの特性:「データ」をどう考えるか.
- とくにデータの「種類」, “質的データ”とはなにか?
- ここで扱う「データ表」の形式と表の相互の関係を知る.
- 対応分析・数量化とは何か. 簡単な例で確認.
- 資料[その1][その2]では, 準備として対応分析・数量化法III類の思想の根底にある, 探索発見的データ解析の要点を述べる.

簡単なデモ, 例1, 例2(JMPスクリプトを利用)

(*)このスライドの最後に添付してある.

この例で, “なにが質的データ”なのか?

対応分析法(CA)の誕生の経緯(抜粋)

- フランスのベンゼクリ(J.-P. Benzécri)の提唱した手法.
- 1962年(頃), 言語解析・語彙用語分析, 方言の分析として応用例を紹介.
- レンヌ(Rennes)で行った講義録(6課程, 1963):
“Statistique et structure des langues naturelles: Essai de synthèse mathématique”. (「自然言語の統計学と構造: 数学的総合化に関する小論」)
- つまり, 非定型・質的データの分析となる.
- B. Cordier(B. Escofier; 故人)が, 理論を整理, プログラムを開発, 1962~1963年頃.

フランス語のアクサン・テギュ, アクサン・グラーヴ, …などは省く

- ベンゼクリが編集主幹の学術誌：“Cahiers de l’Analyse des Données”を通しての研究活動. 自分の主張をこうしたジャーナルで“フランス語”で発表.
- フランス国内で, 独自の研究展開, 普及をみた非常に稀な方法論.
- ベンゼクリは変わった行動をとる人で, 英語圏ではほとんど知られることはなかった. [M.O. Hillの論文(1974)で紹介] 裏話: ベンゼクリはこれを見て批判(激怒)
- 彼の弟子(?)たちが多数いて, 英語での発表などが出回るようになり, この手法の存在が知られるようになった. [カリスマ的・教祖的で, 信奉者が多い]

たとえば, ...

- (故) B. Escofier, L. P. Cazes等による数理の体系化.
- L. Lebart (社会調査, テキスト型データ解析).
- E. Diday (パターン認識, シンボリック・データ解析)
- M. Roux (分類), M. Jambu (データ解析と分類)
- Greenacre (理論の進展)
- 各人ともさまざまな研究, 実務の分野で活躍.
- (故) 林知己夫グループとの接触に始まり, 日仏間の研究者交流も進んだ.

テキストにあげた参考文献を参照のこと.

数量化法III類とは

- “数量化法”は，数量化法III類にかぎらず，体系的に展開された一連の手法．“日本で生まれた稀な方法論”．
- 数量化法III類（パターン分類）は，数量化法の中の1つの手法．1952年頃に提唱された．
- マーケティング・リサーチ，社会調査で実証利用された．
- 数理的には対応分析法に同じだが，アプローチ（定式化），思想は異なる．次第にわかってきた．

(つづき)

- ベンゼクリは「遠い東の国に, (早くに) 自分と同じようなことを考える人がいたとは驚きである」と述べている.
- “数量化法”, “数量化理論”の研究展開については森本(2005)に詳しい.
- このJMRAに“林知己夫ライブラリー”がある.

数量化法・数量化理論(quantification method/theory)と呼称.
ローマ数字で「Ⅰ類, Ⅱ類, Ⅲ類, …」との命名は飽戸弘による. 多次元尺度構成法(MDS)の手法を含めて, (約)「Ⅵ」まである.

対応分析法とは

- 数理的・数学的には“ある行列”の“特異値分解” (S.V.D.)あるいは“固有値問題”を解くことに帰結.
- 数学的に数式で示される事実を, アルゴリズム(算法)としてプログラミングし数値計算処理を行う. 多くは近似計算となることにも注意.
- 数量化法III類も同じと考えてよい.
- “合成変数”(合成指標)の生成の1つと考える.
- この点に注目すると, 数理的には主成分分析(PCA)や次元縮約を伴う判別分析などに類似の操作.

用いる統計ソフトウェアによって, 得られる解がわずかに異なることがある. スコア標準化の有無, 同時布置図の方法などの違い.

(つづき)

- 登場する用語, たとえば「特異値と特異値ベクトル」「固有値と固有ベクトル」「寄与率」「寄与度」など.
- さらに対応分析法では, (ベンゼクリが使った)独自の用語句, 方言が登場する(後述).
- これらが, 対応分析法の利用上, どのような意味を持つかを知ること. 例: 固有値は成分スコアの分散
- 数学的な定式化や証明などは不要. 計算はコンピュータ内のソフトウェアが処理する(誤用・濫用のおそれ).

確認: 特徴をいくつか, ...

- 数量化法Ⅲ類 (quantification method, type III)
 - 林知己夫により提唱された手法 (1952年頃).
 - 多数ある数量化法 (quantification methods) の一つ.
 - 当初は”パターン分類(法)”と呼称 (外的基準のない場合の数量化法の1つ).
 - 当時のコンピュータの性能を考えると, 稀有な発想であり独創的な方法 (コンピュータ処理が難しかった).
 - 質的データに意味ある数値を付与する (尺度化).
 - 「Ⅰ類, Ⅱ類…」というローマ数字の呼称は飽戸弘氏の命名による通称.

(つづき)

- 対応分析法 (Analyse des Correspondances)
 - 正式にはAFC (Analyse Factorielle des Correspondances) という.
 - ベンゼクリにより提唱 (1962～1963年頃), 林の約10年後.
 - CA: Correspondence Analysisとして英語圏に紹介された.
 - 対応分析 (法) と命名 (大隅・林他⇒国内で初めて紹介).
 - 当初は呼称「関連分析」も用いていた.
 - コレスポンデンス分析, コレスポンデンス・アナリシスなど (多分, その本質的な意味が分からなかったのでこんな名称が出た).

analyse factorielle \doteq multivariate analysis (多変量解析) のこと.

- ベンゼクリほかと林知己夫ほかの接触, 日仏研究交流.
- ベンゼクリの招聘要請(飛行機は嫌い, 人には会わない, 奇人でカリスマ的存在).
 - 代わりにM. Roux(ルウ)の招聘, Analyse des Donnéesの紹介
 - データ解析(Analyse des Données)グループとの交流
- 研究集会開催など
 - 国際研究集会への参加, 協力
 - 日仏科学協力セミナーの開催
- L. Lebart(ルバール)他との共同研究など
 - 林ほかの国際意識調査研究
 - “WordMiner”もそうした一環で開発されたソフトウェア(Lebart, 大隅ほかとの共同研究成果から)

- ベンゼクリの用いる“方言”で記述・説明されてきたこと.
- ほとんどがフランス語圏の研究であったため, 的確に理解されてこなかった.
- 用語例をいくつか挙げる. 物理学・力学系とアナロジーがある(ベンゼクリの出自によると思われる).
- これらをこの講座でも用いることになる.

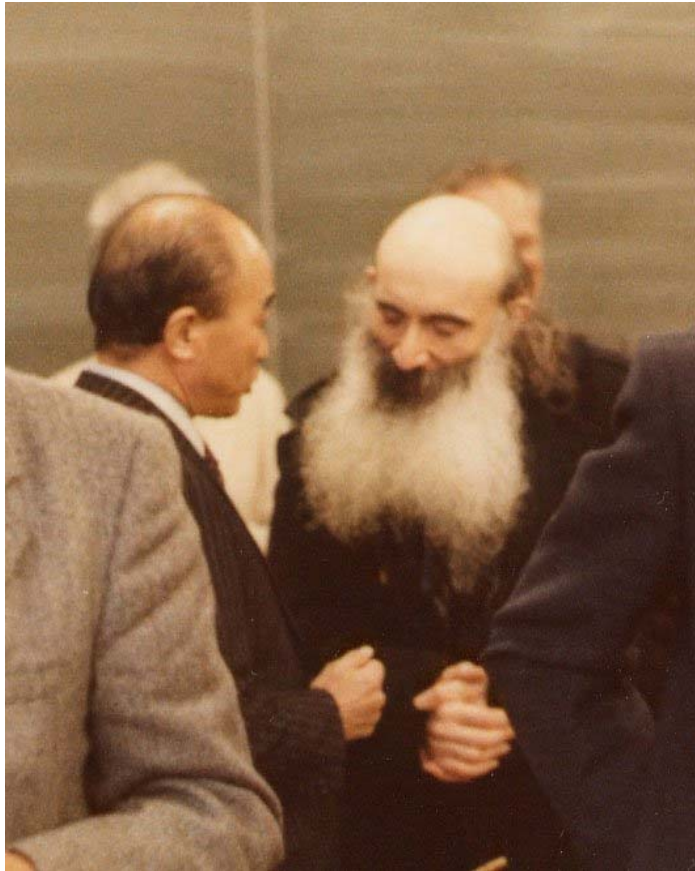
プロフィール, 雲(nuage/英語のcloud), カイ二乗距離, 重心座標系, 質量, 慣性, モダリティ(カテゴリー)またはフォルム, 寄与度(絶対, 相対), 分布の同等性, バート表(多重クロス表), インジケータ行列, 追加処理・追加要素...[案内リーフレットにあるようなキーワード]

- 通常の統計学の用語に“翻訳・読み替え”が必要となる.

- 英語圏の文献も登場し，次第に知られるようになった．テキスト(配付資料)の参考文献を参照．
- 英語圏の書籍を通じて知られるようになったこと．
- いくつかの統計ソフトウェアに実装．たとえば，SAS，JMP，SPSSなど．名称がコレスポンドンス分析となった．
- フランスの統計ソフトウェア^(†)には必ず実装されている．
- 類似の手法が，異なる名称で“さまざまな分野”で登場した．定式化が異なるために，一見すると別の手法のようにみえる．
- 数量化法III類がその典型例で，ある時期まで異なる手法とわれてきた．

(†) おまけ: ソフトウェアをlogiciel(ロジスイエル)という，意図的にこの言葉を用いて母国語を守る．コンピュータもかつて calculateur electroniqueと言ったが，いまはほとんどordinateur(オルディナトゥール)という

★めずらしいスナップショット



故・林知己夫氏＋ベンゼクリ氏



日仏データ解析・データの科学
の研究者交流(初期)

おもな類似手法・関連手法(1)

- 類似手法と関連手法
 - 双対尺度法 (dual scaling; 西里静彦 (1980)).
 - 逆反復平均法・集群分析法 (reciprocal averaging method; M. O. Hill (1973, 1974) 他) [生態学・エコロジーで多用]
 - 等質性分析 (homogeneity analysis; Gifi (1990), J. Meulman (1984) 他)
- とくに逆反復平均法は, 古典的な固有値問題の数値解法の1つである“ベキ法” (power method) に類似.

さまざまな分野で似たようなアプローチが生まれ, 類似の研究展開がある／あったこと.

おもな類似手法・関連手法(2)

- その他の関連手法

- 多重対応分析法(多重クロス表・バート表の対応分析:
MCA: Multiple Correspondence Analysis)
- 対数線形モデルとの関連研究, N. Lauro他(1982),
Hudon(1990), Choulakian(1988)[多数の研究あり]
- 非対称対応分析法(NSCA: non symmetrical CA, N.
Lauro(1994))
- 正準対応分析法(Canonical CA)
- 連関分析法(AA: Association Analysis)とその変形
L. A. Goodman(1986), Agrestiほか
(*) Goodman, Agresti等は, AAはCAに同じと主張
- その他: Subset CA, Joint CAなど(Greenacre, 1984,
2007)ほか

両者に共通した概念・思想

- “データ解析はデータ主導型”である (data driven).
- 単なる“データ処理 (data manipulation, handling) ではない.
- “帰納的”かつ“(仮説)発見的”(探査的)である.
- J. Tukeyの提唱した“探索的データ解析”(EDA)に通底(1960頃から). 実際二人とも, このEDAを重視.
- ベンゼクリは“Analyse des Données”, 林は“データの科学”(data science)を提唱, 通じるものがある.

帰納的(inductive), 発見的・探査的(discovery, exploratory)

EDA: Exploratory Data Analysis

Analyse des Données \neq data analysis (まったく同じではない)

“データの科学”(data science)は, 林ほかが1990年代から使い始めた用語. とくに統計的検定への疑問提起.

★メモ: 通常の分析時の検討要素(言葉だけ)

- 全数調査(全体) vs 標本調査(部分)
- 確率的アプローチ vs 非確率的アプローチ
- 確率標本 vs 非確率標本(便宜的標本など)
- 探索的 vs 確証的
- データ主導 vs モデル主導
- 帰納的 vs 演繹的
- 仮説発見的 vs 仮説検証的
- データの質を重視 vs 誤差は見逃す(緩い対応)

◎分析の戦略次第で結果(内容, 質)は異なる.

データ解析における視点, 観点. どういう立場をとるか.
たがいに対比の関係にある言葉, どう異なり, どう類似するのか.

データ科学の理念に沿って展開

- 理想は確率的アプローチでありたい。現実には、むずかしい。では、どうするか。
- 的確な“データ収集方式”，“調査方式（モード）”の設定がかなめ。ゴミを集めて分析してもゴミ。
- “データは集まるもの”ばかりではない。
- 意図（目的）をもって合理的・的確な方法で“集めるもの”とする精神。
- “帰納的”かつ“仮説発見的”であること。
- データ主導型であること。

データ収集方式 (collection mode), 調査方式 (survey mode)
帰納的 (inductive), 発見的・探索的 (discovery, exploratory)

(つづき)

- つねに記述的であり, 分析的である.
- マイニング(発掘)と探査・発見に通底する.
 - (故)水野欽司によると“手にしたデータは骨までしゃぶりつくす”という精神.
- コンピュータ支援(CA), ソフトウェア利用は不可欠(過信は禁物). 誤用・濫用の回避.
- 多変量解析・多次元データ解析に必須のプラットフォーム.

記述的(descriptive), 分析的(analytic), マイニング(mining)
探査(exploratory), 発見(discovery), コンピュータ支援(CA;
computer assisted)

★メモ: データ収集環境の変化と調査誤差

- 伝統的な“母集団”と“標本”を考えるという構図がなり立たなくなっている(ようにみえる). 要注意.
- 最近の情報通信技術を用いて“(ほぼ)全体・全数”が利用可能だという考え方が登場.
例: POS, 電子カード情報, GPS情報, IoTなど.
例: スマートフォン, SNS上データなどのログ情報.
- “集まるデータ”(集まっているデータ)と“集めるデータ”とは異なること.

(つづき)

- 測定や調査(調べるという操作)をどう考えるか.
- (標本設計に関わる)標本誤差の影響は低減し“非標本誤差”の影響, 重要度が高まったこと.
- 全数の取得が可能になってもなお, 誤差の介入は(おそらくは)避けられない. 例: 回答バイアス, 測定誤差, 無回答誤差など. ⇔全体が見えない?
- “母集団”と“標本抽出”という手続きで得た“標本”を扱う構図を前提としても, しなくても(≡全数利用), “誤差をどう考えるか”は重要な課題.

調査誤差の分類と認識は重要な事項. “総調査誤差”から調べる
ことが重要となっている[大隅他(2011)]

調査誤差とデータの品質は表裏の関係, 下を参考

http://wordminer.org/wp-content/uploads/2013/04/228_0.pdf

(つづき)

- “母集団”と“標本抽出”という手続きで得た“標本”を扱う構図を前提としても、**しなくても**(\equiv 全数利用)、“**誤差をどう考えるか**”は重要な課題.
- 併せて**確率的・統計的事象**をどう考えるかがある.
- 流行りのビッグデータ・アナリティックスでは、これらをどう考えているのだろうか.
- 流行りのエビデンス・ベース(evidence based)にも注意. 多くは“必要条件”しかわからない(可能性).

情報通信技術, ビッグデータ, クラウド, ソーシャル・メディア
総調査誤差, とくに標本誤差と非標本誤差, 偏り(バイアス)

まとめ: データ科学に向けて

- 利用分野の専門的知識と統計的知識の融合的利用が前提. “学際的”から“**横断交流型**”へのシフト.
- コンピュータ利用とその周辺技術(プログラム開発能力)のスキルが必須.
- “**データ収集方式・調査方式**”(mode)の研究の重要性.
- “**データの科学**”の概念の重要性.
- 造語: “**ダータロジー**”を提案した人もあった. [(故)脇本和昌氏]
- “**ニューメラシー**”の重要性.

学際的(inter-disciplinary), 学術横断交流型(cross-disciplinary), データの科学(data science), **ダータロジー**(datalogy), **ニューメラシー**(numeracy; numerate+literacy)
最近では, 質的研究やエスノグラフィーの見直し

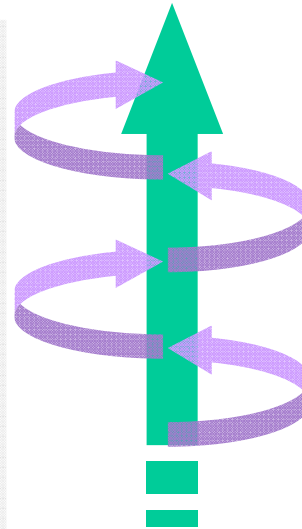
データ処理の効率化と弊害

- 関連領域の学際的な相互利用, 情報の共有の進展
 - 計算機統計学 (computational statistics)
 - グラフィカル表現法 (graphical presentation)
 - 統計ソフトウェア (statistical software)
- PC, 統計ソフトウェアの普及で物事が“暗箱化”した弊害.
- 研究の細分化, 内容・情報がかなり複雑となった.
- 要点は…
 - 集めたデータの履歴・出処を調べよ
 - 統計ソフトウェアにおんぶしてはいけない (自動化の過信?)
 - なんでもできる, と思うな
 - 方法論の“利用の“限界”を知ること
 - なにが, どこまでできるのか, できないのか
 - “べき・べからず”を知ること.
 - 知識要素の連想, 相互の関連づけが重要

◎対応分析の利用時にもすべて通底すること

要約1: 探索的アプローチと確証的アプローチ

- 探索的アプローチ
- (Exploratory)
- 初動探査
- 特徴抽出
- データの視覚化
- (仮説) 発見的
(heuristic, discovery)
- 試行錯誤的



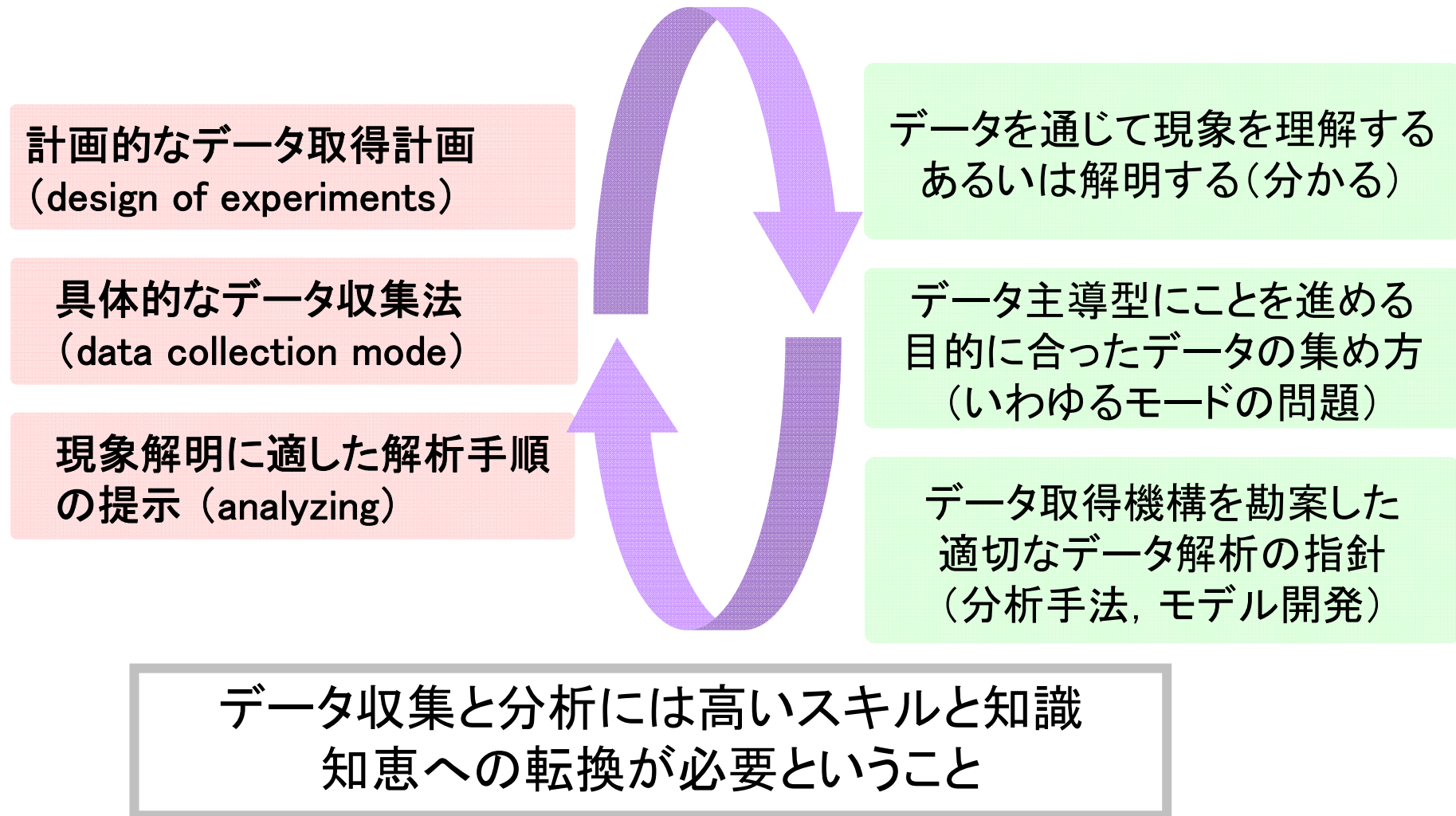
- 確証的アプローチ
- (Confirmatory)
- 統計モデルの構築
- 仮説検証的

> 対応分析・Ⅲ類では根底に…

ループはらせん状に上に向かう(探査→発見→洞察)
問題解決の方向に向かって機能する
対応分析法などはいわゆる**仮説発見的ツール**として有効
いわゆる**デミング・サイクル(PDCA)**を想起

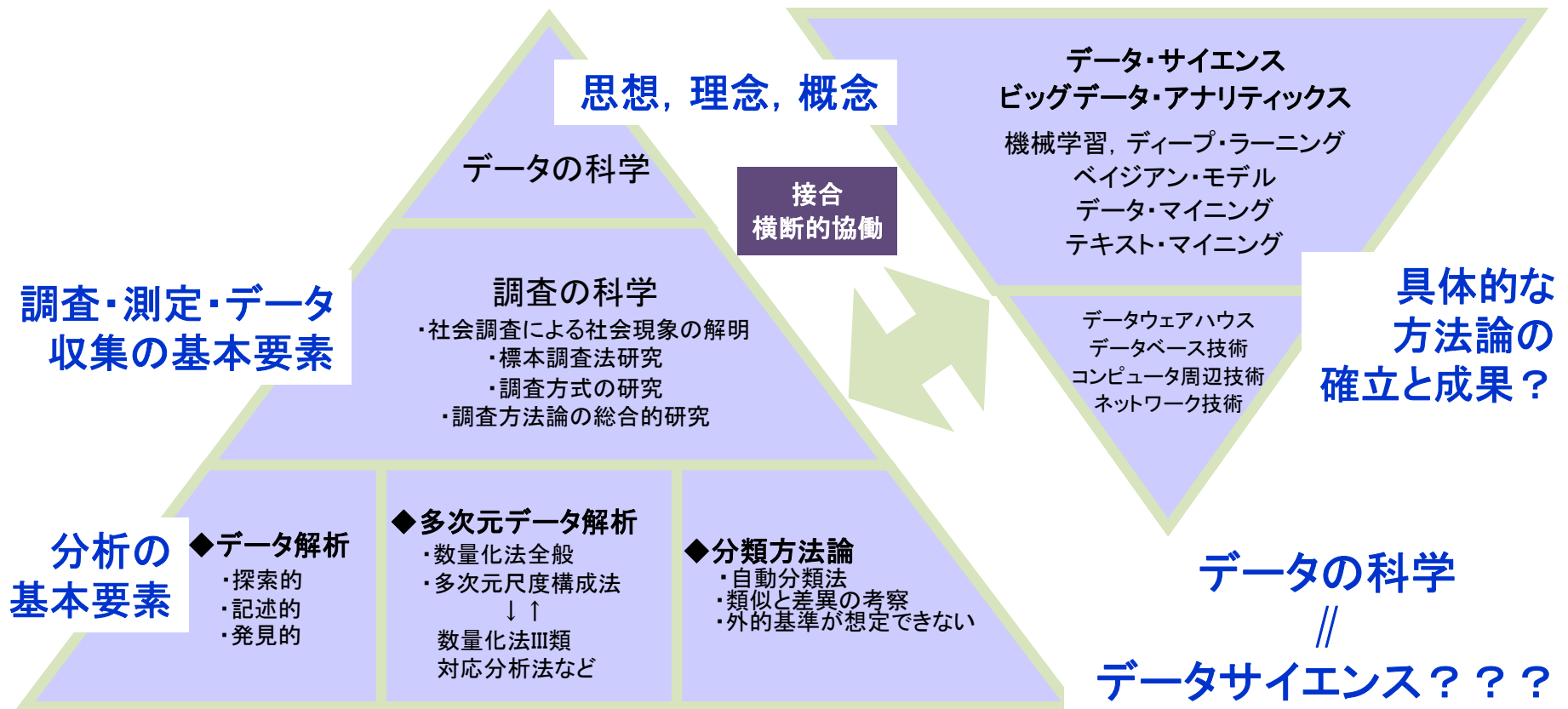
デミング・サイクルまたはデミング・サークルとは
PDCA: plan-do-check/see-action

要約2:「データ科学」の3つの要素(要点のみ)



林知己夫(2001):データの科学, 朝倉書店.

★参考：伝統的な考え方とビッグデータ



“データの科学”のスキーム(概要)[例:林知己夫氏]
最近話題のビッグデータ・アナリティックスはどうか

基本情報:ここから述べること

- データ特性, とくにデータの“種類”を知る
- 基本のデータ表, とくに“2元データ表”について
- “尺度”による分類と質的データ
- 量的データのコーディングと尺度化
- 例による確認をいくつか

データ構造とデータの特徴

- 多変量・多次元データ解析で扱うデータ形式の基本は“多変量構造”の“行列形式”のデータ表.
- つまり2元(two-way)の“行”と“列”からなるデータ表.
- 集計表の“表側”, “表頭”と読み替えてもよい.
- 対応分析法で扱うデータ表も“2元データ表”(two-way data table)の形式を前提.
- 対応分析法が適用できるための若干の要件がある.
- “データの性質”とくに“種類”と“形式”の整理が必要.

データ表の形式は「行列」型だけでないこと
“データ”という表現が曖昧である

対応分析法で扱うデータ表の要件(重要)

- “2元データ表”(two-way data table)であること.
- データ表の各要素(各セル内の値)が“非負の数値”であること. [測定単位が同じ]
- 行または列の“プロフィールが意味のある”データ.
- “プロフィール”, つまりデータ表の行または列の“比率のパターン”が意味を持つデータ表.
- さらに, “ストレッチ・プロフィール”を考えること.
- 行あるいは列に“質的データ”とみなせる“標識”があること. ⇔はじめにみた「旅行データ」

「2元」(two-way)と「多元」(multi-way)は異なる
2元で行列形式を考えることがポイント
プロフィール(profile), 比率・割合のデータ, 質的データ

たとえば, ...

- 典型例が“2元クロス表”あるいは“分割表”.
- (0, 1)型データ行列, インシデント行列. これは2元クロス表の特別な場合. ⇔数量化法III類の基本型
- 多重クロス表(バート表), クロス表の並置型行列.
- アイテム・カテゴリー型, インジケータ行列, 完備排反型行列.
- 多くの統計表(数値が非負の集約データで, 前述の要件を満たすとき).
- 比較的柔軟で, 適用範囲の自由度が高い.

クロス表(cross-classified table), 分割表(contingency table)
インシデント行列(incident matrix), アイテム・カテゴリー型,
インジケータ行列(indicator matrix), 完備排反型(complete
disjunctive form; forme disjonctive complete)

(つづき)

- どんなデータでも適用できるわけではない. 対応分析法に適用可能な要件, つまり“相性”がある.
- 分析に適した“データ表”への加工・変換が必要.
- “データ収集方式”に強く関連する.
- 社会調査: “調査票”や“質問文”の設計に関わること.
- 適切な質問文はうまい分析, 結果につながる.
- これを“構造化する”と言い換えてもよい.
- 以上の確認には“データの特性”とくに“質的データ”とは何かを知ることが必要.

準備:「データ」をどう考えるか？

- 「データの性質」を“3つの要素”から考える.
 - ①データの**種類** (kind, type) ⇔ 分類区分は？
 - ②データ表の**形式** (data table) ⇔ フォーマットは？
 - ③データの**規模** (size, scale) ⇔ 大きさは？
- 細かいことには触れず, 対応分析法の説明に必要な情報に絞って要約する.

別の講座「探索発見的データ解析」における重要課題.

(つづき)

- データの“種類”を分類区分すること.
- とくに“質的データ”とはなにかを知ること.
- 「データ」という言い方は曖昧だが許容.
- ここでは分析対象とする“要素単位”の表象程度に考えておく.
- つまり, 数値だけでなく, 文字, 画像, …広範囲に考える. ⇔ 定性的, 質的である.

データの分類

- データ解析で扱うデータとは？
- ここで「測定値」「観測値」と言うが、数値とは限らない。
- コンピュータ処理上の都合で、文字・記号であっても数値化(このことが誤解を生む)。
- 見方, 用途によって, いくつかに分けられる.
 - 数学的な分類(連続的変量, 離散変量)
 - 尺度による分類(名義, 順序, 区間, 比例)
 - あらたな分類, 別の見方
 - 数値化, 非数値化
 - 構造化, 非構造化, 半構造化
- ここでは“量的データ”と“質的データ”に大別する。
- とくに“質的データ”を考えること。

数学的な分類

連続的変量
あるいは計量的変量
(continuous variable)

長さ, 重さなどを測定したデータは
連続的と考える

実際の測定値は離散的だが,
連続体と考える(実数直線上に対応)

離散的変量
あるいは計数的変量
(discrete variable)

車の台数, 製品個数, 人数など

整数値, 離散的に観測される

- 原則, 実数で, “連続体”か, “離散的”か, と考える.
- 長さ, 重さ, 車の台数, …などは計数的(可算的)である
- 「生活満足感」はどう測る, 測れるのか?

「尺度」による分類(重要!!!)

- 数学的分類だけでは現実の世界, 実際のデータの説明には十分に対応できないことがある.
- 他の分野(心理学, 計量心理学, 社会心理学など)から出てきた見方(現実的, 実用的である).
- “数学的な分類”と“尺度による分類”の関係を知ること.
- とくに“尺度(scale)による分類”が重要である.
- “尺度”とは, たとえば, 態度尺度, 意識尺度のようなものを想定すればよい(ただし, これだけではない).
- “尺度による分類”を, 簡単な例を見ながら考える.

尺度(scale), 態度尺度・意識尺度,
態度尺度構成(attitude scaling)

★参考：態度尺度構成という考え方

- 心理学・計量心理学，社会学，社会調査などでは尺度構成を重視する.
- 態度尺度構成 (attitude scaling) という考え方.
- さまざまな“心理尺度”，“態度尺度”がある.
- 1次元尺度構成の例
 - サーストン尺度，ライカート尺度 (リッカート尺度)
 - ガットマン尺度 ⇔ 数量化法Ⅲ類・対応分析法に発展
 - オスグッドのSD尺度構成 (semantic differential scaling)
- 多次元尺度構成法 (MDS) への展開，発展.
- こうした考え方を念頭に質問文と選択肢を作成してきた.
- データをどう加工処理・演算しているかが重要.

心理尺度として得点化が可能，あるいは尺度化 (スケーリング) が可能との考え方が根底にある.

態度(評定)尺度の簡単な例

例1: 典型的な7段階尺度

+3	+2	+1	±0	-1	-2	-3
非常に 良い	かなり 良い	やや 良い	どちらとも 言えない	やや 悪い	かなり 悪い	非常に 悪い

例2: 4段階尺度(ウェブ調査, ラジオ・ボタン形式)

A. あなたの「普段の生活での気持ち」についておうかがいします。

AQ1 あなたは、現在の生活にどの程度満足していますか。

(あてはまるものを1つ)

- | | |
|--------------------------------|---------------------------------|
| <input type="radio"/> 十分満足している | <input type="radio"/> やや不満である |
| <input type="radio"/> 一応満足している | <input type="radio"/> きわめて不満である |

- いずれの例も, どうスコアリングするのか.
- 単純に加減乗除してよいのか. 平均や標準偏差を算出するが, それが可能な条件は何か. [演算可能性は?]

確認：調査における質問文と選択肢，形式

- 選択肢の構成
 - 単一選択の例，
 - 複数選択の例，複数回答，多岐選択
 - 二項選択，...
- 尺度構成の有無
- 形式（レイアウト）の多様化（とくにウェブ調査）
 - さまざまな形式が使える
 - マルチメディア対応も可，...
- 内容・表現の（微妙的な）違いに注意する
- とくに，「どちらでもない」「ふつう」などの有無
- これらよく考え，“事後の分析利用に適した設計”が肝要.

★参考：ウェブ調査における部品

- 調査票設計の自由度が格段に向上したこと.
- デザイン要素の例
 - ラジオ・ボタン, チェックボックス; テキスト・フィールド, テキスト・ボックス; セレクト・ボックス (プルダウン, ドロップ)
 - グリッド／マトリクス
 - プログレス・インジケータ, アナログ尺度
 - マルチメディア系: 画像 (動画, 静止画), 音声...
- 質問紙型の調査票であっても, 設計の自由度がある.
- “非標本誤差” とくに測定誤差の生起の可能性を考慮.

名義尺度あるいは名目尺度

- 名目的な意味しか持たないデータ (nominal scale).
- 分類尺度, 区分尺度という言い方もある.
- 付与の数値は名目的 (nominal) に与えた識別符号にすぎない.
- 与えた数値 (コード) の大きさや差異には意味はない.
- 原則として四則演算 (加減乗除) がむずかしい.
- 数値表記しても, 計算 (加減乗除) ができるとは限らない.
- 多くの場合 “割合 (比率, %)” に定量化して扱う.

ライカート尺度構成 (Likert scaling), 態度尺度構成
(*) 日本ではリッカートと呼ぶことが多いようだ.

★参考：尺度の序列化・極化

- 選択肢の単極化型と両極化型（尺度の向きを考えること）
- 例：「満足度」, 「賛否」の尺度化
 - 例1: 単極型尺度: 「非常に満足」「かなり満足」「あまり満足でない」「まったく満足でない」と序列化したとき
 - 例2: 両極化型尺度: 「非常に満足」「かなり満足」「やや不満」「まったく不満」と, 対比させたとき
 - 例3: 「賛成する, 賛成しない」と「賛成, 反対」は異なる.
- 例1は「順序尺度」, 例2は「名義尺度」, 例3は, 前者が順序尺度, 後者は名義尺度に相当.

単極化型 (unipolar scale), 両極化型 (bipolar scale)

(つづき)

- 自記式調査(例: 郵送調査, ウェブ調査)では, 調査票のレイアウト, デザインなどが回答者行動に影響する.
- とくに, “測定誤差, 無回答誤差”などが看過できない.
- つまり, “非標本誤差”の発生・影響が無視できない.
- 一般に, 質問文や選択肢を作るとき, ワーディングに十分な考慮が必要.
- 基本は, “回答者が違和感や抵抗なくすみやかに”回答可能な質問文とすること.
- これは調査データの質に関係すること.

調査誤差の分類が重要

標本誤差・非標本誤差, 測定誤差, 無回答誤差

★メモ:よく知られた回答行動の例

- 社会的望ましさ (social desirability)
 - 回答者が、本来あるべき回答とは別の回答がより好ましいと考えて回答を提供する傾向のこと。たとえば、「微妙な質問」の場合などに起こる。
- 黙従傾向 (acquiescence)
 - 回答者が、質問文の記述内容にたいし、その文脈に関係なく、同意する傾向があるという性質。いわゆる「(なんでも)“はい”と回答し易い傾向」(yes tendency)のこと。
- 初頭効果 (primacy effects)
 - 選択肢型質問で、回答者が回答選択肢の一覧から前のほうにある回答選択肢を選びやすいという傾向のこと。

(つづき)

- 新近性効果 (recency effect)
 - 回答者が、選択肢型質問のリスト内にある最後の選択肢を優先して選ぶという現象のこと。
- 微妙な質問 (sensitive question)
 - 微妙な内容に関わる情報あるいはそうした話題を扱う質問文のこと。微妙な内容とは、回答者が回答をためらう、あるいは正確な回答を避けるような場合。違法薬物使用経験、犯罪履歴など。
- 労働最小化行動 (satisficing behavior)
 - 回答者が調査質問を読むことや回答するための労力を割かないこと。提示された個々の回答選択肢をきちんと評価せずに、回答者が適当だと思った選択肢が見つかった時点でそれを選んでしまう、自分に都合よくはしよるということ。
 - 労働最小化行動、最小限化行動、最小限回答行動などさまざまな訳語がある。

(つづき)

- ストレートライニング(直線的回答傾向)
 - グリッド形式(マトリクス形式), それに類似したレイアウトの質問群の中の, すべての単一選択質問にたいして同じ回答を選ぶ傾向があること.
 - 同じ列にあるすべての回答選択肢を選ぶ傾向. それに近い回答選択をする傾向. 労働最小化行動の1つ.
- 回答の偏り(バイアス)
- 回答拒否や調査不能, 無回答, 回答中断などの生起.

◎回答者の回答行動を測る, あるいは評価するさまざまな概念や指標がある. 知っておくと便利だろう.

順序尺度あるいは順序尺度

- 名目尺度であって、選択肢の用語間に意味としての「差異」または「順序関係」がある場合 (ordinal scale)
- この場合も原則として“加減乗除ができない”。
- 名義尺度、順序尺度の計量化の1つの手続きが割合・比率を求める操作。そして、CAやMCAがある。

例1:

「満足」、「あまり満足でない」「満足でない」⇔順序尺度
「かなり満足」「満足」「やや不満」「不満」⇔名義尺度

例2:

順序尺度と名義尺度の混在

ウェブ調査の調査票からの例をいくつか

ライカートは別の見方, スコアリング

選択肢に意味あるとして, 平均・標準偏差など統計量を算出する

ウェブ調査の例

Q17 ひとくちにおいて、あなたは今の生活に満足していますか、それとも不満がありますか。
(ひとつだけ)

- ☐ 1. 満足
- ☐ 2. やや満足
- ☐ 3. やや不満
- ☐ 4. 不満

名義尺度
順序尺度ではない

ここで、「紙に印刷された書籍(紙の本)」についてうかがいます。

Q12 「今後の技術の進歩にともない、“紙の書籍”の時代が終わり、次第に“電子書籍”の時代になるだろう」という意見があります。
あなたはこれについて、どのようにお考えですか。
(ひとつだけ)

- ☐ 1. 非常にそう思う
- ☐ 2. そう思う
- ☐ 3. あまりそうは思わない
- ☐ 4. まったくそうは思わない
- ☐ 5. いずれもが共存すると思う
- ☐ 6. わからない・考えたことはない

尺度の混在

どちらも「加減乗除」は？

区間尺度あるいは間隔尺度

- 数値の差(差分)が意味を持つ場合を区間尺度あるいは間隔尺度(interval scale)
- 測定の単位がある.
 - 身長や体重, 気温, ネジの径などは測定の単位がある(kg, g, cm, mm, °Cなど)
 - (測定)精度に関係することに注意
- 測定単位が変わることは変動に影響することに注意.
 - 分散の大きさ, 変化に関係する; 分布の変動(形)が変わる
- 個々の観測値の数値の差が個々の測定対象の差異に対応.
 - 身長に関して言えば、165cmの人は170cmの人より5cm低い
 - 気温に関しては、35°Cは25°Cよりも10°C高い
- 四則演算(加減乗除)が可能. 注: 除算は比例尺度のみ

比例尺度あるいは比率尺度

- 区間尺度であって、原点(ゼロ)を基点とする**比が意味を持つ場合、比率尺度・比尺度)ともいう**(ratio scale)
 - 身長や体重およびネジの径などはデータの大小関係(差)以外に比も意味を持つ
 - Aの体重が45kg、Bの体重が90kgであればBはAの2倍の体重である
- 区間尺度であって**比例尺度でない典型例**
 - 温度の場合は -10°C と 25°C が何倍違うかというような比較ができないので区間尺度であっても比例尺度ではない。
- “変動係数”は比例尺度にのみに適用可能。

変動係数(C.V.: coefficient variation)は「標準偏差」/「平均」($\times 100$)。

区間尺度あるいは間隔尺度の例

- ① 身長と体重を測定, 記録する(たとえば170cm, 65kg)
- ② ネジの径(mm)をマイクロメータで測定し、数値を記録
- ③ 毎日の気温を摂氏(°C)で測定
- ④ 駅の改札を単位時間内に通過する人数を測定
- ⑤ 塗装面のピンホール数の測定

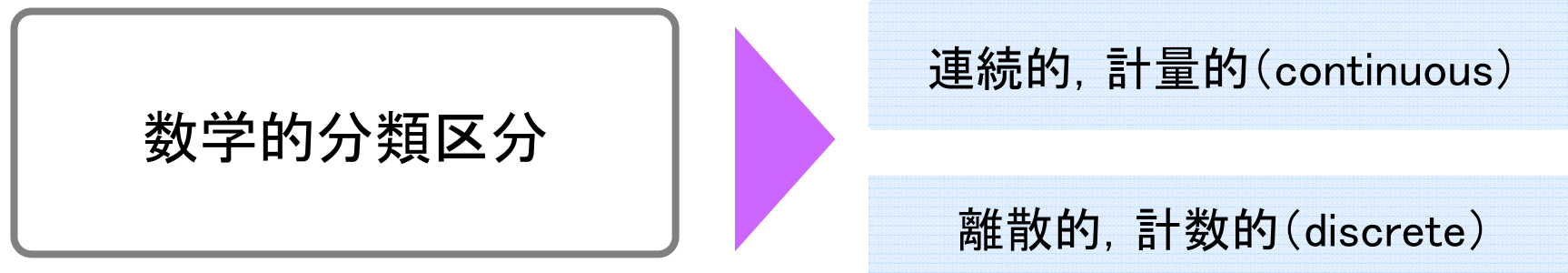
- 数学的分類に従うと, ①～③は連続的, ④～⑤は離散的.
- 加減乗除が可能
- ただし③は原則加減のみ＝単に区間尺度ということ.

以上の考え方に従うと, ...

- 多くのデータの解釈が“**実感**”に合う.
- 最近の統計ソフトウェアは, これを**ユーザ指定により認識**できる. SAS/JMPほか
- 指定した条件の合った分析が選べる(誤った分析回避).
- 統計量の算出に統計的知識が反映されるようになった(多少は誤用が防げる).
- 対応分析法・数量化法III類との関係では, 量的データ(区間尺度・比例尺度のデータ)をカテゴリー化して名義尺度化, 順序尺度化して用いるような場面が重要(後述).

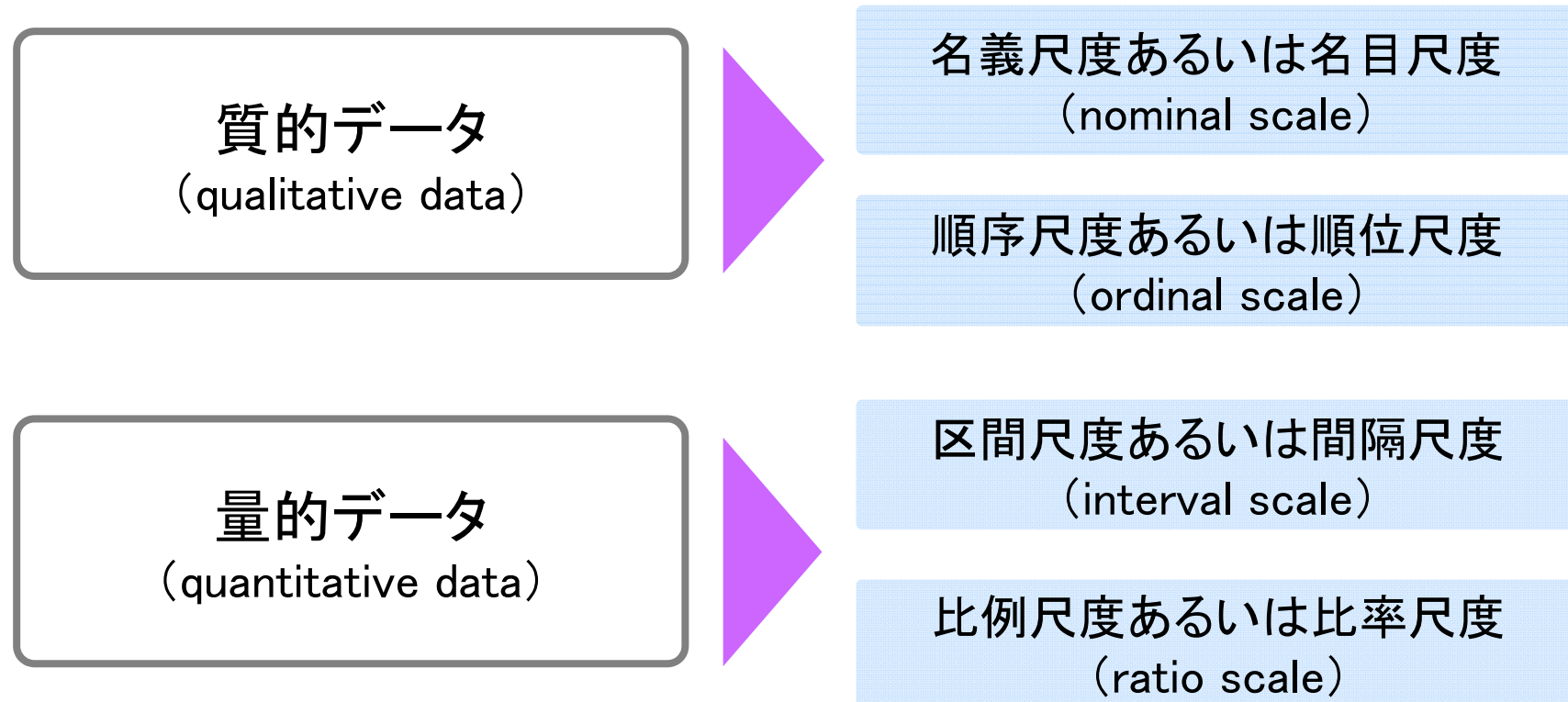
以上を, まとめると, ...

要約1: 数学的分類



統計的議論, 確率分布 (連続分布, 離散分布) や確率変数 (連続型変数, 離散型変数) など, 統計推論を進めるうえで便利な分類

要約2: 尺度による分類



「定量的」「定性的」という言い方もある
データ解析を行ううえではこうした見方が便利
しかし、さまざまな意見・異論がある

注意として, ...

- データの分類(構造を知る)とは, 操作上の“一つの目安”.
- データの特性をどう“分類区分する”かと, それを何らかの“加工”を経て, 別の情報に置き換えること(情報の変換)とは分けて考える.
- 加工を経ることによって“情報の質と内容”が変わる.
- 質的データは情報が少なく, 量的データが多いとはならない (そう単純ではない)情報量は, その時々のお作によって変わる.
- 多くの統計解析では, 数量に変換して計量的に考察する. それが難しいことがある. 対応分析法は1つの策.

(つづき)

- 変換操作による情報の質的, 量的な変化・損失がある.
- とくにテキスト型データの扱いは配慮が必要.
- どのような分類を行っても, 状況に応じて解釈や分析上の操作に都合のよい形で用いる.
- “当たり前”のように思えるが, 調査票・質問文と選択肢の設計時に重要なこと.
- 調査誤差, とくに“非標本誤差”(測定誤差, 無回答誤差)に影響する.

エディティング, コーディングの影響

- とくに(量的データの)カテゴリー化, 区分化は難問の1つ.
- (最適な区分化は)おそらく, いまでも明確な解はない.
- 単純に数値化が可能とは限らない. よって“数量化”の思想が重要となる.
- テキスト型データを含む“文字情報の分析”もそうした課題の1つ.

例1: ポストコーディング(日本でいうアフターコーディング)

例2: 辞書編集, 同義語・類語の編集, コーパス作成など

- 対応分析の適用上, この操作が重要となることがある.
- 簡単な例を3つほど, ながめる.

エディティング, コーディング(プレコーディング, ポストコーディング)

いくつかの簡単な例をみる

例1:「血圧」を測ること.

- 実測単位のまま(量的データ)を分析に用いるか,「非常に高い」「高い」「低い」「非常に低い」と名義尺度化(質的データ)するか.

例2:調査で「世帯所得収入」を問う.

- 質問文にテキスト・フィールドを設けて収入金額を書き入れてもらう(量的データ)か,選択肢型の順序尺度とするか.

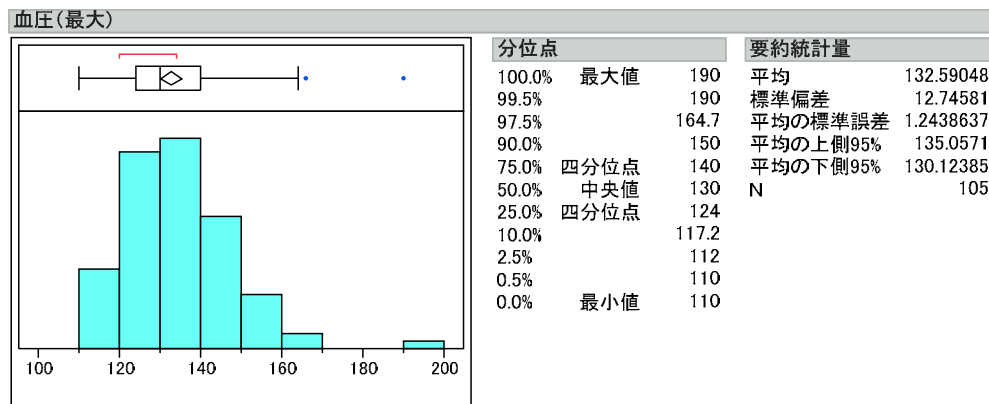
例3:ウェブ調査の回答所要時間の探査.

- 非常に偏った裾の長い分布となるが,それを直接分析に利用するには問題がある.では,どうカテゴリー化するか.

所得・収入などの実金額を問う方式は回答に偏りを生じるおそれや,無回答が多くなる可能性がある(好ましくない問い方)(回答の偏り)
カテゴリー化,区分化は難題の1つ(最適な区分化とは,への解は?)

例1:「血圧」を測ること

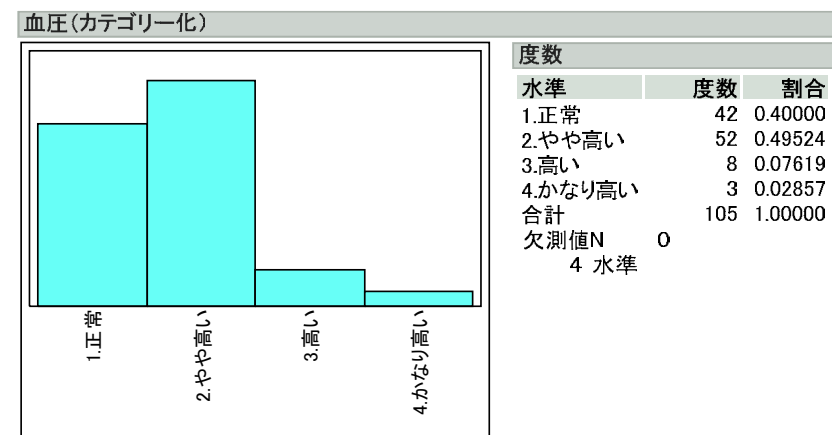
- 実測の単位そのままを分析に用いるとき(量的データ)
- 血圧が「正常」「やや高い」「高い」「かなり高い」と順序尺度化したとき(質的データ, 順序尺度化)
- どちらが良いかはそのときの状況による(分析目的など).



実測データのヒストグラムと記述的統計量

血圧(最大) < 130 ⇒「正常」
130 ≤ 血圧(最大) < 150 ⇒「やや高い」
150 ≤ 血圧(最大) < 160 ⇒「高い」
血圧(最大) ≥ 160 ⇒「かなり高い」

たとえば, 4区分にカテゴリー化



例2: 世帯所得収入の区分化

- 直接, テキスト・フィールドを設けて収入金額を書き入れてもらう(量的データ).

[注: 測定方式としてはやや疑問. 偏りや無回答・拒否の増加]

- 何段階かの選択肢に分け順序尺度のように扱う(質的データ). [選択肢の作り方が問題]
- 記入で得た回収データをどう使うか, で“選択肢の区分”の影響がある(調査票設計の影響, 回答の偏り, 測定誤差).

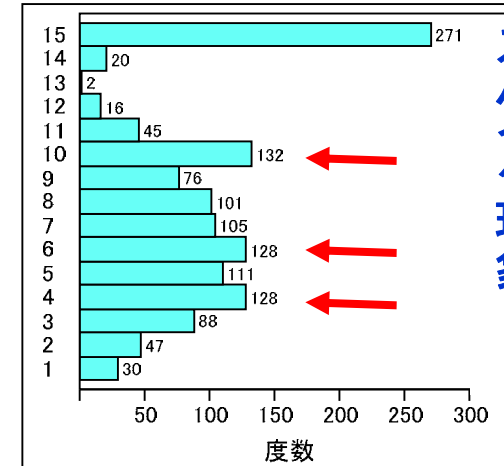
大変立ち入ったことで恐縮ですが、お宅(世帯)全体の年収(税込み)
(あてはまるものを1つ)

- | | |
|---|--|
| <input type="radio"/> 200万円未満 | <input type="radio"/> 200～300万円未満 |
| <input type="radio"/> 300～400万円未満 | <input type="radio"/> <u>400～500万円未満</u> |
| <input type="radio"/> 500～600万円未満 | <input type="radio"/> <u>600～700万円未満</u> |
| <input type="radio"/> 700～800万円未満 | <input type="radio"/> <u>800～900万円未満</u> |
| <input type="radio"/> 900～1,000万円未満 | <input type="radio"/> <u>1,000～1,200万円未満</u> |
| <input type="radio"/> 1,200～1,500万円未満 | <input type="radio"/> 1,500～1,800万円未満 |
| <input type="radio"/> 1,800～2,000万円未満 | <input type="radio"/> 2,000万円以上 |
| <input type="radio"/> <u>わからない・答えたくない</u> | |

(このレイアウトは問題)

一変量の分布

問8-6: 世帯年収



区分は正しいか?
スパイク現象

66

sumi

例3: ウェブ調査「回答所要時間」分布

- ウェブ調査では“回答所要時間”の探査が重要である.
- 非公募型のウェブ・パネル(部分的に確率的パネル)を利用したウェブ調査で得た1つの例を示す.
- “パラデータ”(プロセス・データ)を用いて分析を行う.
- 同じような場面がウェブの“ログ解析”で生じるだろう.
- 回答所要時間分布は, 裾の長い典型的な“ロング・テイル”かつ“単峰分布”となる.
- そのまま量的データとして, 分析に使えないことが多い.
- カテゴリー化, 質的データへの変換で情報の質が変わる.

ウェブ調査とウェブ・パネル, 非公募型パネル
パラデータ(paradata), 裾の長い(重い)分布
ロング・テイルの分布(ジフ分布, パレート分布)

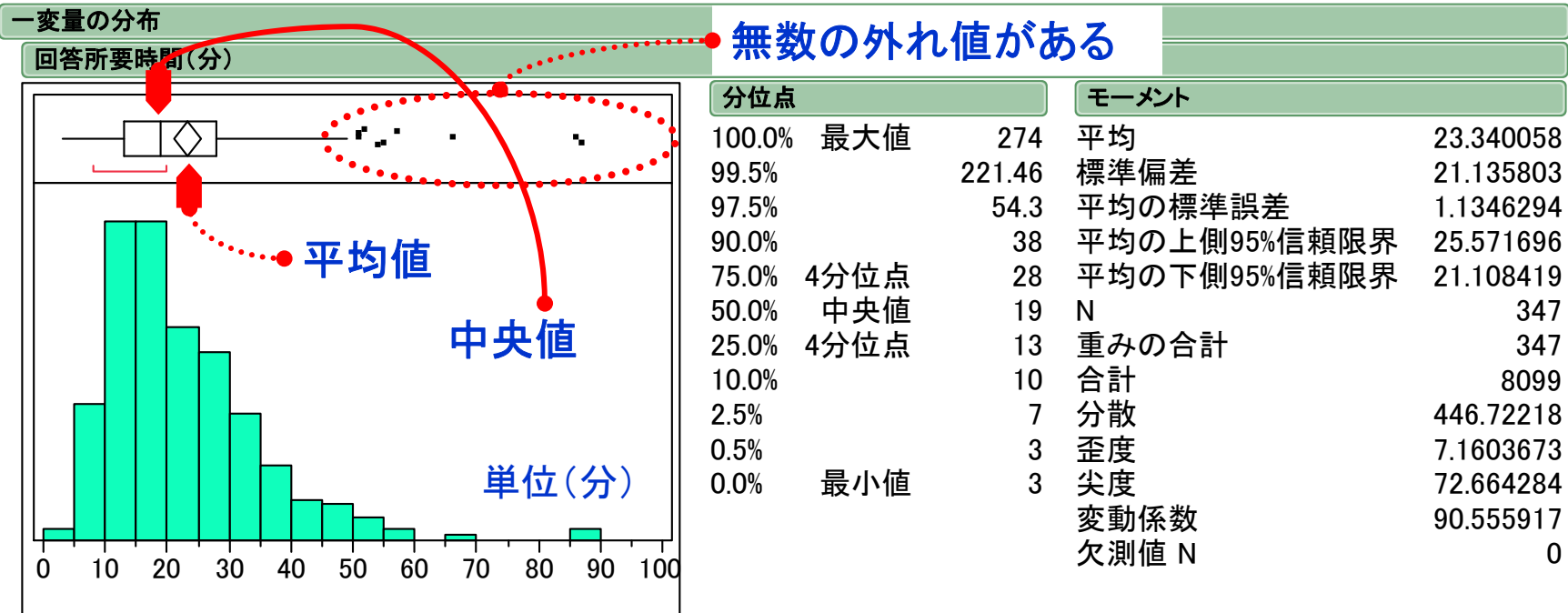
調査の概要

- テーマ:「情報に関する調査」(実験調査)
- 調査方式:ウェブ調査
- 実施期間: 2011年09月09日 17:00 ~ 2011年09月13日 09:00まで
- ウェブ・パネル:非公募型パネル(部分的に確率的パネル)
- 予想回答所要時間:約20分
- 計画標本の大きさ:766(人)[男性(412), 女性(354)]
- 回収標本の大きさ:347(人)[男性(175), 女性(172)]
- 有効回収率:45.3(%) [“参加率”というほうがよい]

回収標本の年齢分布(人口統計学的変数の一部情報)

	サンプル数	15～19歳	20～24歳	25～29歳	30～34歳	35～39歳	40～44歳	45～49歳	50～54歳	55～59歳	60～64歳	65～69歳	この中にはない
合計	347	22	25	32	39	39	44	25	26	29	39	27	-
		6.3	7.2	9.2	11.2	11.2	12.7	7.2	7.5	8.4	11.2	7.8	-

回答所要時間の分布の探査



非常に裾の長い単峰分布となる(右端カットした). ロングテイルの典型例.

記述的統計量の特徴を読む.

中央値=19(分)<平均値=23.3(分), この数分の違いは大きい

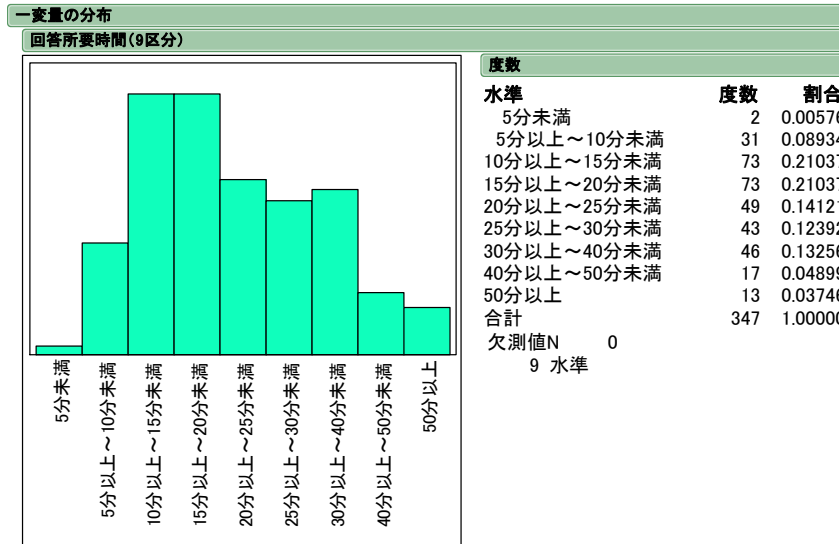
標準偏差=21.14(分)

歪度や尖度は? [これがゼロのときが正規分布]

変動係数に注目(91%もある), その他は?

データの確認

回答所要時間のカテゴリー化の例



9区分にカテゴリー化

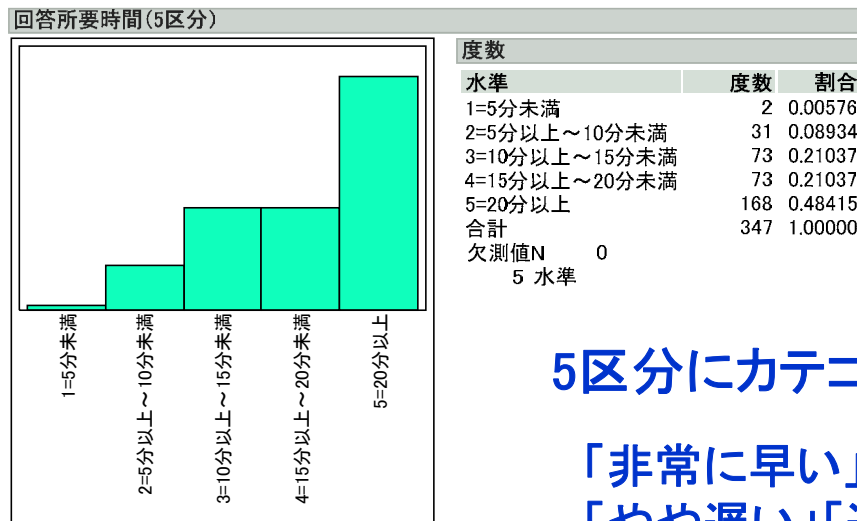
こうした区分化操作で分析結果が左右される

どうするか, 試行錯誤あるいは“外部情報源”の利用

どちらを使うかで, 分析・解釈が異なる可能性.

1つの目安は(裾を調整し)
“元の分布型”に近づける.

いわゆる「打ち切り型」の場合は要注意



5区分にカテゴリー化

「非常に早い」「早い」「やや早い」
「やや遅い」「遅い」「非常に遅い」としたら?

なぜ、スコア化や区分化を行うか

◎スコアリング(評点化)について

- 質的データの場合、ライカート尺度のようなスコアリングの考え方が浸透.
- なるべく尺度点の中央に対称に近い分布となるように考えたいこと.
- のちの分析処理での計量化, 演算処理可能性を想定.
- (頭の中に)なるべく“正規分布”的になってほしい, という見方があること.

(つづき)

◎カテゴリー化・区分化について

- 量的データのカテゴリー化では、なるべく“元の分布”の特徴を保持したいという意図がある.
- 出来るだけ、変換後の尺度感がそろそろ(等間隔的)であってほしいという見方. 実際はそうならないことが多い.
- 分位数(四分位数, 八分位数, 十分位数など)を形式的に当てはめるという方法もある.
- 区分点をどうするかが重要なこと. たとえば, 年齢区分や所得区分などは, 形式的に行うことは要注意.

◎うまく表現できないが, およそ上のような考え方を, 分析時に念頭におくこと(経験則, 知識と合わせて).

別の分類もあるだろう

数值的か非数值的か



数值的 (numerical)
数量, 数値, 計数として表記されるもの

非数值的 (non-numerical)
文字, 記号, イメージ (静止画, 動画),
音声など

(例: 探査衛星データ, YouTubeなど)

(つづき)

構造的か非構造的か

最近, 注目の
キーワード

構造的データ(structured data)
カテゴリー化, タグ化, コード化などを
行いデータベース化など整備されたデータ

(例: POSデータ, 顧客DBなど)

非構造的データ(unstructured data)
とくに何も措置されない裸のままのデータ
(一般には扱いにくい)
(ソーシャルメディアで扱うデータなど)

(ソーシャルメディアで扱うデータなど)

半構造的データ(semi-structured)
自由回答質問など

(集める, 集められるデータ)

一方向発信
回答制御できない
(集まるデータ)

双方向的(インタラクティブ)設計で調整・制御可能な面

© Noboru Ohsumi

データの種類による組合せ

- さまざまな組合せが考えられる.
- 量的データと質的データから考える.
- 変量に観察する向き(主と従)があるとき.
 - 例: 独立変数と従属変数
- すくなくとも”4通り“がある.
 - (量的データ) × (量的データ), (量的データ) × (質的データ)
 - (質的データ) × (質的データ), (質的データ) × (量的データ)
- 扱う“変量数”の個数(1変量, 2変量, 多変量).
- “カテゴリカル・データ分析”としてさまざまな研究が展開.

組み合わせの例とツール

	量的データ×量的データ		質的データ×質的データ	
	用いる統計量	グラフィカル表現	用いる統計量	グラフィカル表現
1変量	基本統計量 (平均値, 分散, 標準偏差など)	度数分布 ヒストグラム (周辺分布)	割合・比率など	単純集計・項目別集計 (周辺分布)
2変量	相関係数	散布図 (同時分布)	連関性の測度 例:ピアソンのカイ 二乗統計量	モザイク図ほか (同時分布の観察)
多変量	相関係数行列	散布図行列 (多変量連関図) 連関図	連関性測度の行列	連関図 多重クロス表 (バート表)の図化
適用手法の例	主成分分析, 因子分析など (*) 因子分析は使い方に注意		数量化III類, 対応分析法, 多重対応分析法 対数線型モデルなど	

右側に注目する

ここから述べること

- 調査事例による(2元)クロス表の観察.
- クロス表, 分割表の関係を「連関性の測度」で測る.
- 連関性の測度は無数にある.
- とくに, 対応分析で重要な“(ピアソンの)カイ二乗統計量”の扱い方.
- さまざまな“(2元)データ表”の生成.
- データ表の相互の関連性を知ること.

「環境意識調査データ」のクロス表探査

- この例で、意識調査データの“記述的分析”を試みる.
- 環境意識調査「都市環境の住みやすさに関する調査」, 林・水野・大隅他(1983~1985)(3年間の継続調査)
- 調査概要:
 - 調査計画, 標本設計・サンプリングは研究者(我々)が実施
 - 標本抽出枠: 選挙人名簿を利用
 - 実査は調査機関 (輿論科学協会)に委託
 - 共通の調査票を用い「6地域で年度を変えて」実施
 - 個々の実施回で“共通の標本抽出法”を適用(確率標本).
 - 調査方式: 調査員による訪問留置・自記式

標本抽出枠・枠(サンプリング・フレーム), 確率標本,
調査方式(調査モード)

6地域の調査内容の概略

調査年次	調査回 および対象地域	計画標本数 (人)	回収標本数 (人)	回収率(%)
1983年 (S58年)	千里ニュータウン	1,800	1,205	67.0
	千葉市市街部	1,440	768	53.3
1984年 (S59年)	箕面市の一部	870	588	67.6
	千葉市市街隣接部	1,440	1,074	74.6
1985年 (S60年)	東京都江東区	1,170	755	61.4
	三鷹市	900	635	64.5
全 体		7,620	5,030	66.0

注1: 調査時点が異なることに注意(やや非等質である)

注2: 各地域の住民を代表している(代表性はほぼあるとみてよい)

注3: “調査地域”は一つの層別変数と考えられる(⇔名義尺度)

注4: ここでは, 主に1983年の回収標本を用いる

初動探査:クロス表分析の要点

- “記述的統計量”による観察
 - 度数(頻度)の観察.
 - 行和, 列和などの表示
 - 割合の算出, 全体%, 行%, 列%など.
 - 割合の検定を行うこともあるだろう.
- “分析的統計量”の算出と観察
 - 連関性の測度の算出
 - “カイ二乗統計量”の算出と“統計的検定”の操作
- 状況に応じてデータ加工処理
 - 選択肢のプーリング
 - はずれ値(outlier)のチェック
 - 無回答, DK, などの調整(除外など)
- グラフィカル表現法による探査
 - モザイク図ほか(視覚化)

対応分析に入る前の
事前分析・初動探査

ここでの目標は, ...

- まず, 2変量(項目)の関連を“クロス表”で調べる.
 - クロス表の観察, 相関の有無.
 - モザイク図, 分析的統計量の利用
- クロス表の“周辺分布”と“同時分布”を知る.
- 連関性の測度として“ピアソンのカイ二乗統計量”を用いる.
- さらに多数の項目(多変量)の観察.
- “多重クロス表”(バート表), モザイク図で探査する.

- 1項目, 2項目と多数項目
- クロス表あるいは分割表, その周辺分布, 同時分布
- 連関性の測度, ピアソンのカイ二乗統計量
- 多重クロス表(バート表)

2項目の関係:クロス表とモザイク図

- 多数の質問から次の2つの質問項目 I , J を取り出す.
- 用いた質問と選択肢を以下に示す.
- いずれも質的データ(名目尺度・順序尺度)である.

質問 I : あなたは、いま住んでいるまちが気に入っていますか.

1. たいへん気に入っている
2. まあ気に入っている
3. あまり気に入っていない
4. まったく気に入っていない

質問 J : あなたの住んでいる地区は、都市としては緑(みどり)が多いと感じますか. それとも少ないと感じますか.

1. かなり多い
2. 多いほうである
3. ふつう
4. 少ないほうだ
5. きわめて少ない

質問(questions), 選択肢(options, choices)

調査で得た元の多変量構造のデータ表

確認

★1983年(S58年)調査データ_subset - JMP

ファイル(F) 編集(E) テーブル(T) 行(R) 列(C) 実験計画(DOE)(D) 分析(A) グラフ(G) ツール(Q) アドイン(N) 表示(V)

質問(変量)

★1983年(S58年)調査データ... ノート /Users/Ohsumi_Noboru/Desktop

列(20/0)

- 地点番号
- サンプル番号
- Q1(1): 1: 年号コード
- Q1(1): 1: 年号コード_ラベル
- Q1(1): 2: 昭和のとき, ○○年
- Q1(2): まちは気に入っているか
- Q1(2): まちは気に入っているか_ラベル
- Q2(1): 緑が多いか_ラベル
- Q9(3): 今後も住みたい_ラベル
- Q10(1): この10年でよくなったか
- Q2(1): 緑が多いか
- Q9(3): 今後も住みたい
- Q10(1): この10年でよくなったか
- F(1): 性別
- F(1): 性別_ラベル
- F(2): 年齢
- 年齢区分

行

- すべての行 1,973
- 選択されている行 27
- 除外されている行 27
- 表示しない行 0
- ラベルのついた行 0

	Q1(2): まちは気に入っているか	Q1(2): まちは気に入っているか_ラベル	Q2(1): 緑が多いか_ラベル	Q9(3): 今後も住みたい_ラベル	Q10(1): この10年でよくなったか_ラベル	Q2(1): 緑が多いか
1 56	1	たいへん気に入っている	多い方である	当分はここに住みたい	変わらない	
2 53	2	まあ気に入っている	多い方である	当分はここに住みたい	変わらない	
3 42	2	まあ気に入っている	多い方である	当分はここに住みたい	変わらない	
4 56	1	たいへん気に入っている	かなり多い	当分はここに住みたい	変わらない	
5 5	2	まあ気に入っている	多い方である	当分はここに住みたい	ややわるくなった	
6 4	2	まあ気に入っている	多い方である	当分はここに住みたい	変わらない	
7 5	2	まあ気に入っている	多い方である	当分はここに住みたい	ややよくなった	
8 4	2	まあ気に入っている	多い方である	ずっとここに住みたい	ややよくなった	
9 4	2	まあ気に入っている	多い方である	当分はここに住みたい	ややよくなった	
10 5	2	まあ気に入っている	かなり多い	当分はここに住みたい	ややよくなった	
11 5	1	たいへん気に入っている	かなり多い	ずっとここに住みたい	非常によくなった	
12 4	2	まあ気に入っている	ふつう	当分はここに住みたい	非常にわるくなった	
13 5	2	まあ気に入っている	かなり多い	当分はここに住みたい	ややよくなった	
14 5	1	たいへん気に入っている	かなり多い	当分はここに住みたい	変わらない	
15 5	2	まあ気に入っている	多い方である	できれば転居したい	ややわるくなった	
16 4	1	たいへん気に入っている	かなり多い	当分はここに住みたい	変わらない	
17 5	2	まあ気に入っている	かなり多い	当分はここに住みたい	ややわるくなった	
18 4	1	たいへん気に入っている	多い方である	ずっとここに住みたい	変わらない	
19 4	2	まあ気に入っている	かなり多い	ずっとここに住みたい	変わらない	
20 4	2	まあ気に入っている	多い方である	ずっとここに住みたい	ややよくなった	
21 5	2	まあ気に入っている	多い方である	ずっとここに住みたい	ややよくなった	
22 39	1	たいへん気に入っている	多い方である	ずっとここに住みたい	ややよくなった	
23 54	3	あまり気に入っていない	ふつう	できれば転居したい	非常にわるくなった	
24 54	2	まあ気に入っている	多い方である	当分はここに住みたい	変わらない	
25 37	1	たいへん気に入っている	多い方である	当分はここに住みたい	変わらない	
26 55	2	まあ気に入っている	少ない方だ	当分はここに住みたい	ややよくなった	
27 37	1	たいへん気に入っている	多い方である	ずっとここに住みたい	変わらない	
28 56	2	まあ気に入っている	多い方である	できれば転居したい	ややよくなった	
29 37	2	まあ気に入っている	多い方である	当分はここに住みたい	変わらない	

回答者(個体・サンプル)

(質問I) × (質問J)の
クロス表を生成する

質問I

質問J

この質問の特徴

- 質問Iでは4つの選択肢が, 質問Jでは5つの選択肢がある.
- 2つの質問の選択肢による区分に従い, 調査の全回答者を分類集計する. その要約表が2元クロス表である.
- 以下のような追加情報が生じる(ポストコーディング).
 - 無回答(NA: No Answer)や回答拒否
 - 「わからない」(DK: Don't Know)という回答もある
 - このコードを事後に加え $5 \times 6 = 30$ のマス(セル: cell)のクロス表

選択肢を, オプション, カテゴリー(分類区分のこと)
対応分析では, モダリティ(modalité)という
2元クロス表(two-way cross-classified tables)

2元クロス表の確認

- クロス表の内部セルの分布＝質問Iと質問Jとの同時分布
- 質問Iの和：質問Iの周辺分布となる（ここでは行和）
- 質問Jの和：質問Jの周辺分布となる（ここでは列和）

質問 I 質問 J	1.大変気に入っている	2.まあ気に入っている	3.あまり気に入っていない	4.気に入っていない	行和
1.かなり多い	166	131	6	2	305
2.多いほう	239	598	40	2	879
3.ふつう	86	324	55	0	465
4.少ないほう	26	146		5	228
5.少ない	7	36	20	6	69
列和	524	1235	172	15	1946

(質問 I, J の同時分布)

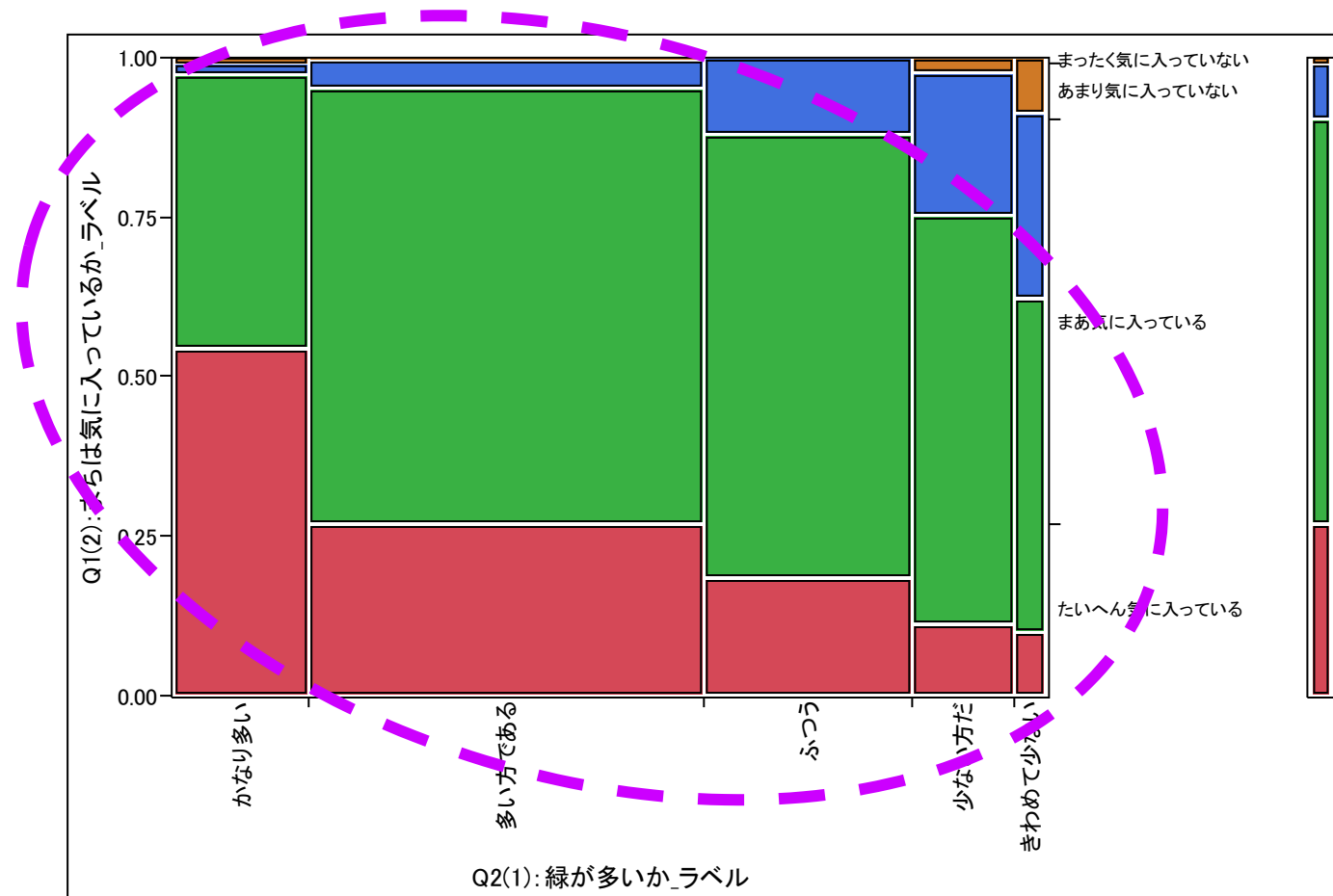
(質問 I の周辺分布)

(質問 J の周辺分布)

注：ここでは「無回答」は除外した。

周辺分布, 同時分布

モザイク図による傾向の観察(JMPの出力)



相関があるようだ, それをどう測るか

データの確認

2項目の関係を「モザイク図」で観察

- 2つの質問, “まちが気に入っている”と“緑が多いか”は関連・相関がありそう.
- このことは, 回答者の“意識として”そうだということ.
- 実際の“緑地度”と重ね合わせるとわずかな相関がある(約0.4程度).
- おおまかには“住めば都”という傾向と読める.
- 人口統計学的変数(属性)や, 他の選択肢質問が関連するのではないか.
- 実施年度, 対象地域, 性別などでも差違があるだろう.

ここで, “緑地度”の分布という“外部情報源”(external information)があることに注目. これを“(優れた)基準”として利用できるかもしれない. 調査設計で事前に考えておくこと.

2項目の関係をクロス表と統計量で観察

Q2(1):緑が多いか_ラベルとQ1(2):まちは気に入っているか_ラベルの分割表に対する分析

分割表

Q2(1):緑が多いか_ラベル	Q1(2):まちは気に入っているか_ラベル				
	たいへん気に入っている	まあ気に入っている	あまり気に入っていない	まったく気に入っていない	
度数					
全体%					
列%					
行%					
かなり多い	166 8.53 31.68 54.43	131 6.73 10.61 42.95	6 0.31 3.49 1.97	2 0.10 13.33 0.66	305 15.67
多い方である	239 12.28 45.61 27.19	598 30.73 48.42 68.03	40 2.06 23.26 4.55	2 0.10 13.33 0.23	879 45.17
ふつう	86 4.42 16.41 18.49	324 16.65 26.23 69.68	55 2.83 31.98 11.83	0 0.00 0.00 0.00	465 23.90
少ない方だ	26 1.34 4.96 11.40	146 7.50 11.82 64.04	51 2.62 29.65 22.37	5 0.26 33.33 2.19	228 11.72
きわめて少ない	7 0.36 1.34 10.14	36 1.85 2.91 52.17	20 1.03 11.63 28.99	6 0.31 40.00 8.70	69 3.55
	524 26.93	1235 63.46	172 8.84	15 0.77	1946

各セル内に表示の
情報の説明

注:ここでは「無回答」を除外した.

ここで「行和」「列和」「総和」を100とした
割合を出力に注意. ⇔確率行列(後述)

種々の連関性の測度の観察

Q2(1): 緑が多いか_ラベルとQ1(2): まちは気に入っているか_ラベルの分割表に対する分析

関連の指標				
指標	値	標準誤差	下側95%	上側95%
ガンマ	0.4861	0.0289	0.4295	0.5427
Kendallのタウ-b	0.3035	0.0193	0.2657	0.3413
Stuartのタウ-c	0.2433	0.0163	0.2113	0.2753
SomersのD (C R)	0.2609	0.0171	0.2275	0.2944
SomersのD (R C)	0.3530	0.0223	0.3094	0.3967
非対称ラムダ(C R)	0.0492	0.0236	0.0029	0.0955
非対称ラムダ(R C)	0.0178	0.0094	0.0000	0.0363
対称ラムダ	0.0304	0.0110	0.0087	0.0520
不確実性係数(C R)	0.0812	0.0094	0.0627	0.0996
不確実性係数(R C)	0.0533	0.0063	0.0409	0.0657
不確実性係数(対称)	0.0643	0.0076	0.0495	0.0791

いろいろな連関性の測度がある
それぞれが何かを測っているが
ここでは触れない
※ユーザの要請で追加

Q2(1): 緑が多いか_ラベルとQ1(2): まちは気に入っているか_ラベルの分割表に対する分析

検定			
	N	自由度	(-1)*対数尤度 R2乗(U)
	1946	12	141.17134 0.0812
検定	カイ2乗	p値(Prob>ChiSq)	
尤度比	282.343	<.0001*	
Pearson	340.309	<.0001*	

カイ二乗統計量の表示
(2つの質問の関連を測る指標)
帰無仮説: 2つの質問は独立を検定

(ピアソンのカイ二乗統計量)

$$\chi_p^2 = 340.309$$

★連関性の測度とは？

- 質的データを集約化した情報の項目間の連関性の程度を数値として測る必要がある.
- 2元データ表とくに2元クロス表に要約の2項目間の連関性を測るさまざまな指標がある.
- これらを総称して“連関性の測度”という. [JMPでは”関連の指標“と表記]
- クロス表の非対称性などを測る指標もある.
- とくに“ピアソンのカイニ乗統計量” (χ_p^2) にもとづくさまざまな指標がある.
- 対応分析法ではこの“カイニ乗統計量”が重要な役割.

連関性の測度 (measures of association), ピアソンのカイニ乗統計量

★主な連関性の測度(この他, 無数にある)

クロス表の型	名称と特徴
2 × 2クロス表	ピアソンの χ^2 統計量 (Pearson) [2 × 2クロス表のときは相関係数の二乗に比例]
	イエーツの係数 (Yates), マクネマーの係数 (McNemar), フィッシャーの直接確率法 (Fisher) [超幾何分布が関係する]
$m \times n$ クロス表	ピアソンの χ^2 統計量 [対応分析法と密接に関連]
	ピアソンの χ^2 統計量に依拠するさまざまな指標 ϕ^2 係数 (平均平方関連係数) クラメール係数 (Cramer), ピアソン連関係数 (コンティンジェンシー) ケンドール=スチュアート係数 (Kendall-Stuart)
	選択肢に順序関係がある場合 (順序尺度) の関連性測度 グッドマン=クルスカルの係数 (Goodman-Kruskal), ケンドール係数 (Kendall), ソマーズ係数 (Somers)

注1: ピアソンのカイ二乗統計量から派生する指標が多い

注2: この他, 尤度比カイ二乗統計量がある

とくにピアソンのカイ二乗統計量に注目

- 一例として“**カイ二乗統計量**”による「独立性の検定」を考える.
- 独立性の検定を行う.
 - 帰無仮説「**2つの質問 I, J は互いに独立**」を検定
 - 検定統計量として“ピアソンのカイ二乗統計量”を使う.
 - 一種の分布間の距離を測る指標.
 - (クロス表の寸法が同じとして)これが大きいほど有意となる.
- 対応分析法ではこのカイ二乗統計量という検定統計量がきわめて重要.
- あとで詳しく検討する. スライド資料[その2]とした.

検定結果の解釈

- 形式的に, このクロス表から得たピアソンのカイ二乗統計量による検定結果を読み取る.
- ピアソンのカイ二乗統計量 = 340.309 となって, 有意確率 (p 値) は < 0.0001 であり, 高度に有意である, となる.
- 解釈: 2つの質問の関係は「何らかの関係がありそうだ」となる.
- 仮説「独立である」を棄却しただけであるから, 断定的に「関係がある」とは言えない(かなり確度が高いが).
- 検定特有の背理法的(二重否定的)に, “関係がないとはいえない”(ありそう)という表現.

いわゆる「背理法」のように, 仮説設定と検定を行う.

(つづき)

- しかしこの2つの質問は, おそらくは“関係があるだろう”と想定して設けたはずである.
- 「関係がありそう」という結論だけでは不十分ではないか.
- さらに, 2つの質問文の選択肢にはどういう関係があるのだろうか.
- これへの1つの解を対応分析法が提供する.
- 対応分析法の説明に入る前に, まず, 対応分析法で扱う“データ表の形式”について述べる.
- そして, ピアソンのカイ二乗統計量とそれにもとづく“独立性の検定”について簡単に触れる.

どのような「データ表」を扱うのか？

- “2元クロス表”は典型的な2元データ表である.
- クロス表の形式でなくても2元データ表はさまざま場面で登場する.
- 多変量構造のデータ表からも, いろいろな2元データ表が生成される.
- コーディング処理やデータ表の加工で得られる典型的な2元データ表を調べる.
- とくに対応分析法が適用可能な2元データ表の要件はかなり緩やかなものである.
- つまり, 扱えるデータ表の条件を緩める.

再確認: 対応分析法で扱う2元データ表

- ① “2元(two-way)の行列”形式となっていること.
- ② 各要素(セル)内の数値は“非負の値”であること.
- ③ 行あるいは列の“比率のパターン”, つまり“プロフィール”を考える意味があるような場合.
- ④ あるいはそれに相当する場面を想定できる行列形式の2元データ表.

ある種の多変量構造のデータ行列

- この条件を満たすデータ表は身近に沢山みられる.
- 対応分析法では, この中でもとくにいくつかの形式のデータ表間の数理的な関係が調べられている.
- 数量化法III類との類似性や相似性がある.
- 対応分析法のほうが扱えるデータ表のバリエーションが多い.
- (一般の多変量構造で)なるべく“測定単位がそろって”いて“行・列の比率パターンが意味がある”ような例
 - 例1: 都道府県別に要約の民力データ(負の値を含まない, かつなるべく単位がそろっている変量)
 - 例2: 多くの官庁統計情報の書式

(つづき)

- 「あり・なし」「はい・いいえ」型の応答行列(インシデント行列).
- 典型的な例が「社会調査」で扱うような質的データからなるさまざまなデータ表.
- データ加工(再コーディング: recoding)によるデータ変換とそれで得られるデータ表.
- 既述のように量的データから生成の区分化データも含む.
- 順位データ(ranking data)にも適用される.

テキスト(I 部), 24ページあたりから

例1:「銘柄」を選ぶ(2値応答データ)

サンプル	サンプルが選んだ銘柄
サンプル1	銘柄A, 銘柄C
サンプル2	銘柄B
サンプル3	銘柄A
サンプル4	銘柄B, 銘柄C



サンプル	銘柄A	銘柄B	銘柄C
サンプル1	1	0	1
サンプル2	0	1	0
サンプル3	1	0	0
サンプル4	0	1	1

- 想定場面:「サンプル」(回答者)が「好みの銘柄」を選ぶ.
- 左の表(テキスト型データ)から右の表(コード)に変換.
- (0, 1)データに変換. インシデント行列という.
- 数量化法III類でかならず登場するデータ表形式.
- クロス表の特別な場合と考えてよい.
- 対応分析法で分析・説明できる.

注:ここで,「好き」か「そうではない」か,で「好き,嫌い」とはなっていない.

例2: 好きな清涼飲料を選ぶ(2値応答データ)

回答者 番号	1	2	3	4	5	6	7	8	回答者 番号	1	2	3	4	5	6	7	8
1	1	0	0	0	1	1	0	1	16	0	0	0	0	1	1	0	0
2	1	0	0	0	1	0	0	0	17	0	1	0	0	0	1	0	0
3	1	0	0	0	1	0	0	0	18	1	1	0	0	1	0	0	0
4	0	1	0	1	0	0	1	0	19	1	0	0	0	0	0	0	1
5	1	0	0	0	1	0	0	0	20	1	1	1	0	1	0	0	0
6	1	0	0	0	1	1	0	0	21	1	0	0	0	1	0	0	0
7	0	1	1	1	0	0	1	0	22	1	0	0	0	1	0	0	0
8	1	1	0	0	1	1	0	1	23	0	1	0	1	0	0	1	0
9	1	1	0	0	0	1	1	1	24	1	1	0	0	1	0	0	0
10	1	0	0	0	1	0	0	1	25	0	1	1	1	0	0	0	0
11	1	0	0	0	1	1	0	0	26	0	1	0	1	0	0	1	0
12	0	1	0	0	0	0	1	0	27	0	1	0	0	0	0	1	0
13	0	0	1	1	0	1	0	1	28	1	0	0	0	0	1	0	1
14	1	0	0	0	0	1	0	0	29	1	0	0	0	0	1	0	0
15	0	1	1	0	0	0	1	0	30	0	1	1	0	0	0	1	0

(清涼飲料水の銘柄リスト)

1	2	3	4	5	6	7	8
コカコーラ	ダイエットコーク	ダイエットペプシ	ダイエット7アップ°	ペプシ	スプライト	Tab	7アップ°

100

これをテキスト型データで表すと, ...

回答者番号	回答者が選んだ「好む」清涼飲料	回答者番号	回答者が選んだ「好む」清涼飲料
1	ココーラ, ペプシコーラ, スプライト, 7アップ	16	ペプシコーラ, スプライト
2	ココーラ, ペプシコーラ	17	ダイトコーク, スプライト
3	ココーラ, ペプシコーラ	18	ココーラ, ダイトコーク, ペプシコーラ
4	ダイトコーク, ダイト7アップ, Tab	19	ココーラ, 7アップ
5	ココーラ, ペプシコーラ	20	ココーラ, ダイトコーク, ダイトペプシ, ペプシコーラ
6	ココーラ, ペプシコーラ, スプライト	21	ココーラ, ペプシコーラ
7	ダイトコーク, ダイトペプシ, ダイト7アップ, Tab	22	ココーラ, ペプシコーラ
8	ココーラ, ダイトコーク, ペプシコーラ, スプライト, 7アップ	23	ダイトコーク, ダイト7アップ, Tab
9	ココーラ, ダイトコーク, スプライト, Tab, 7アップ	24	ココーラ, ダイトコーク, ペプシコーラ
10	ココーラ, ペプシコーラ, 7アップ	25	ダイトコーク, ダイトペプシ, ダイト7アップ
11	ココーラ, ペプシコーラ, スプライト	26	ダイトコーク, ダイト7アップ, Tab
12	ダイトコーク, Tab	27	ダイトコーク, Tab
13	ダイトペプシ, ダイト7アップ, スプライト, 7アップ	28	ココーラ, スプライト, 7アップ
14	ココーラ, スプライト	29	ココーラ, スプライト
15	ダイトコーク, ダイトペプシ, Tab	30	ダイトコーク, ダイトペプシ, Tab

これは質的データである

調査で「好きな飲み物の名前をあげよ(自由記述)」と読み替えてみる

データ表の関係を知らること(重要)

- ここで, データ表(=行列)に“便宜的に名前”を付けて整理する.
- 多変量構造の(2元の)データ行列[X表]
- “2元クロス表”[F表]. これが基本となる.
- 質問文の選択肢のコーディングで得たデータ表[C表]
- アイテム・カテゴリー型行列, インジケータ行列, 完備排斥型行列(以下で, “インジケータ行列”とする)[A表]
- “多重クロス表”(バート表, バート行列)[B表]

アイテム・カテゴリー型, インジケータ行列, 多重クロス表(バート表)
バート表とは, 提唱者のC. Burtの名前を付けたもの.

(つづき)

- ここで“多元クロス表”との違いに注意する.
- コーディングの意味(事象の「あり, なし」), 尺度化の意味を考えること.
- 調査データを例として一連の関係を確認する.
- とくに“(2元)データ表”の, 表側と表頭にある標識は何か, 表の各セル内の数値の単位と意味は何か, に注意すること.
- 数式による説明は資料にある. また, あとで述べる.

多元クロス表(multi-way tables), 多重クロス表(multiple table)

ある市民調査の例で確認

- ある自治体で行った市民意識調査の例.
- 市内にある「農業公園についての意識調査」
- 計画標本の大きさ: 1,008人(男性493人, 女性515人)
- 回収標本の大きさ: 411人^(†)(男性178人, 女性233人)
回収率: 41%
- 調査対象者: 市内に居住の成人
- 選挙人名簿から2段無作為抽出, 確率標本
- 調査方式: 郵送調査, 自記式方式
- 市民はこの調査課題に関心なく(当時), 回収率は低い.

テキスト, 26ページあたりから
(†)性別無回答が2名, よって「413(人)」

調査データの確認

(つづき)

- 調査票の質問文から、つぎの2つの質問を選ぶ.

質問I: 昔からの習慣をよく守っているか.

1. 守っている
2. まあ守っている
3. あまり守っていない
4. 守っていない

質問J: 神社や、お寺詣りをよくするか.

1. お寺詣りをよくする
2. たまにお寺詣りをする
3. あまりお寺詣りをしない
4. お寺詣りをしない

注: ここで、「お詣り」は通常は「お参り」だろうが、この調査実施者が用いた質問文のまま表記する.

市民意識調査データの一部(X表)

回収サンプル 番号	地域コード	計画サンプル 番号	この公園 構想を 知ってい ましたか。	この公園構想を 知っていましたか。(選択肢)	この公園構想を 知っていましたか。(選択肢)	1.近くの緑地や公園等 をよく散策している。 (選択肢)	3.昔からの習慣をよく 守っている。(選択肢)	6.神社や、お寺詣りをよくす る。(選択肢)	8.自分のなすべき役割は積極 的に果している。(選択肢)
1	11	181	1	はい	知っていた	あまり散策しない	まあ守っている	お寺詣りをしない	役割はあまり果たしていない
2	11	185	2	いいえ	知らなかった	あまり散策しない	あまり守っていない	たまにお寺詣りをする	まあ役割は果たしている
3	11	188	1	はい	知っていた	散策しない	守っている	あまりお寺詣りをしない	役割はあまり果たしていない
4	11	189	2	いいえ	知らなかった	散策しない	まあ守っている	お寺詣りをしない	まあ役割は果たしている
5	11	198	2	いいえ	知らなかった	あまり散策しない	まあ守っている	たまにお寺詣りをする	まあ役割は果たしている
6	12	199	2	いいえ	知らなかった	あまり散策しない	まあ守っている	たまにお寺詣りをする	まあ役割は果たしている
7	12			いいえ	知らなかった	あまり散策しない	まあ守っている	あまりお寺詣りをしない	役割は果たしている
8	12			いいえ	知らなかった	まあ散策している	あまり守っていない	お寺詣りをしない	役割は果たしている
9	12			はい	知っていた	まあ散策している	まあ守っている	たまにお寺詣りをする	まあ役割は果たしている
10	12			はい	知っていた	よく散策している	まあ守っている	たまにお寺詣りをする	まあ役割は果たしている
11	12	211	2	いいえ	知らなかった	散策しない	無回答	お寺詣りをしない	まあ役割は果たしている
12	12	212	2	いいえ	知らなかった	あまり散策しない	まあ守っている	あまりお寺詣りをしない	役割は果たしている
13	12	215	2	いいえ	知らなかった	散策しない	まあ守っている	お寺詣りをよくする	役割は果たしている
14	13	217	2	いいえ	知らなかった	まあ散策している	守っている	お寺詣りをよくする	まあ役割は果たしている
15	13	221	2	いいえ	知らなかった	あまり散策しない	まあ守っている	あまりお寺詣りをしない	役割は果たしている
16	13	223	1	はい	知っていた	まあ散策している	守っていない	あまりお寺詣りをしない	まあ役割は果たしている
					知っていた	あまり散策しない	まあ守っている	お寺詣りをしない	まあ役割は果たしている
					知らなかった	散策しない	守っている	お寺詣りをしない	まあ役割は果たしている
					知らなかった	まあ散策している	まあ守っている	あまりお寺詣りをしない	役割は果たしている
					知らなかった	散策しない	まあ守っている	あまりお寺詣りをしない	まあ役割は果たしている
					知っていた	よく散策している	守っていない	あまりお寺詣りをしない	まあ役割は果たしている
22	14	238	2	いいえ	知らなかった	あまり散策しない	まあ守っている	たまにお寺詣りをする	役割は果たしている
23	14	239	2	いいえ	知らなかった	散策しない	まあ守っている	お寺詣りをしない	役割は果たしている
24	14	242	1	はい	知っていた	無回答	守っている	お寺詣りをよくする	役割は果たしている
25	14	244	2	いいえ	知らなかった	まあ散策している	まあ守っている	あまりお寺詣りをしない	役割は果たしている
26	14	248	2	いいえ	知らなかった	あまり散策しない	まあ守っている	あまりお寺詣りをしない	まあ役割は果たしている
27	14	251	2	いいえ	知らなかった	まあ散策している	あまり守っていない	あまりお寺詣りをしない	役割はあまり果たしていない
28	15	253	1	はい	知っていた	散策しない	まあ守っている	たまにお寺詣りをする	役割は果たしている
29	15	254	2	いいえ	知らなかった	散策しない	守っていない	お寺詣りをしない	役割は果たしていない

多変量構造のデータ表(X表)
寸法が「サンプル数×項目数」

この多変量構造のデータ表から、2つの質問(2列)を選ぶ

コーディングを行う(C表の生成)

3_昔からの習慣をよく守っている。(選択肢)	6_神社や、お寺詣りをよくする。(選択肢)
まあ守っている	お寺詣りをしない
あまり守っていない	たまにお寺詣りをする
守っている	あまりお寺詣りをしない
まあ守っている	お寺詣りをしない
まあ守っている	たまにお寺詣りをする
まあ守っている	たまにお寺詣りをする
まあ守っている	お寺詣りをしない
あまり守っていない	お寺詣りをしない
まあ守っている	お寺詣りをする
まあ守っている	たまにお寺詣りをする
無回答	お寺詣りをしない
まあ守っている	あまりお寺詣りをしない
まあ守っている	お寺詣りをよくする
守っている	お寺詣りをよくする
まあ守っている	あまりお寺詣りをしない
守っていない	あまりお寺詣りをしない
まあ守っている	お寺詣りをしない
守っている	お寺詣りをしない
まあ守っている	あまりお寺詣りをしない
まあ守っている	あまりお寺詣りをしない
守っていない	あまりお寺詣りをしない
まあ守っている	たまにお寺詣りをする
まあ守っている	お寺詣りをしない
守っている	お寺詣りをよくする
まあ守っている	あまりお寺詣りをしない
まあ守っている	あまりお寺詣りをしない
あまり守っていない	あまりお寺詣りをしない
まあ守っている	たまにお寺詣りをする
守っていない	お寺詣りをしない

X表から

3_昔からの習慣をよく守っている。	6_神社や、お寺詣りをよくする。
2	4
3	2
1	3
2	4
2	2
2	2
2	3
2	4
2	2
5	4
2	3
2	1
1	1
2	3
4	3
2	4
1	4
2	3
2	3
4	3
2	2
	4
	1
	3
	3
3	3
2	2
4	4

C表

寸法が
「サンプル数×2」

- コードブックに従って、コード化したということ
- 印字が見にくいので配布の「テキスト」も参照. 27～29pあたり
- データ表の寸法の変化に注意する.

インジケータ行列への変換(A表の生成)

3_昔からの習慣をよく守っている。	6_神社や、お寺詣りをよくする。
2	4
3	2
<div style="border: 2px solid blue; padding: 10px; display: inline-block;"> C表 </div>	
2	3
3	4
2	2
2	2
2	2
5	4
2	3
2	1
1	1
2	3
4	3
2	4
1	4
2	3
2	3
4	3
<div style="font-size: 2em; font-weight: bold; color: blue;"> 寸法が 「サンプル数×2」 </div>	
3	3
2	2
4	

寸法が
「サンプル数×2」

1.守っている	2.まあ守っている	3.あまり守っていない	4.守っていない	5.無回答	1.お寺請りをよくする	2.たまにお寺請りをする	3.あまりお寺請りをしていない	4.お寺請りをしない	5.無回答	
	1	1				1		1		
1	1		A表				1	1		
	1						1		1	
	1						1			
	1	1						1		
	1					1		1		
	1			1		1			1	
	1						1			
1	1				1		1			
	1		1				1			
1	1						1	1		
	1						1			
	1		1				1			
	1						1			
1	1							1		
	1						1			
	1						1			
	1						1			
	1						1			
	1						1			
	1	1					1			
	1						1			
	1					1				
		1					1			
			1					1		

寸法が「サンプル数×10」

コーディング・データ(C表)からインジケータ行列(A表)へ
インジケータ行列の行和が質問数=2となることに注意
同じく、列和は2つの質問の周辺分布(周辺度数)となること

2元クロス表(F表の生成)

		質問 J					
質問 I	選択肢	お寺詣りをよくする	たまにお寺詣りをする	あまりお寺詣りをしていない	お寺詣りをしていない	無回答	計(行和)
	守っている	41	26	22	15	2	106
	まあ守っている	25	67	45	30	0	167
	あまり守っていない	6	13	34	31	0	84
	守っていない	1	6	7	27	0	41
	無回答	1	4	1	2	7	15
	計(列和)	74	116	109	105	9	413

「質問 I × 質問 J」の2元クロス表
 ここでは、寸法が「5 × 5」のクロス表

2項目の多重クロス表・バート表(B表)

表 24 多重クロス表(バート表) $B = A^t A$ の例(表 23 から生成)

質問	質問	質問 I					質問 J				
	選択肢	守っている	まあ守っている	あまり守っていない	守っていない	無回答	お寺詣りをよくする	たまにお寺詣りをする	あまりお寺詣りをしていない	お寺詣りをしない	無回答
質問 I	守っている	106	0	0	0	0	41	26	22	15	2
	まあ守っている	0	167	0	0	0	25	67	45	30	0
	あまり守っていない	0	0	84	0	0	6	13	34	31	0
	守っていない	0	0	0	41	0	1	6	7	27	0
	無回答	0	0	0	0	15	1	4	1	2	7
質問 J	お寺詣りをよくする	41	25	6	1	1	74	0	0	0	0
	たまにお寺詣りをする	26	67	13	4	4	0	116	0	0	0
	あまりお寺詣りをしていない	22	45	34	7	1	0	0	109	0	0
	お寺詣りをしない	15	30	31	27	2	0	0	0	105	0
	無回答	2	0	0	0	7	0	0	0	0	9

B表

(バート表: $B = A^t A$ から得られる行列; ここで, A^t は A の転置行列)

クロス表の“並置”になっている(2元データ表)

“対称行列”である(寸法が 10×10 ; 2つの質問の選択肢数の和)

インジケータ行列との関係に注意しよう(次ページ)

バート表(B表)とインジケータ行列(A表)

表 24 多重クロス表(バート表) $B = A^t A$ の例(表 23 から生成)

$B = A^t A \Rightarrow$

質問	選択肢	質問 I					質問 J				
		守っている	まあ守っている	あまり守っていない	守っていない	無回答	お寺詣りをよくする	たまにお寺詣りをする	あまりお寺詣りをしない	お寺詣りをしない	無回答
質問 I	守っている	106	0	0	0	0	41	26	22	15	2
	まあ守っている	0	167	0	0	0	25	67	45	30	0
	あまり守っていない	0	0	84	0	0	6	13	34	31	0
	守っていない	0	0	0	41	0	1	6	7	27	0
	無回答	0	0	0	0	15	1	4	1	2	7
質問 J	お寺詣りをよくする	41	25	6	1	1	74	0	0	0	0
	たまにお寺詣りをする	26	67	13	6	4	0	116	0	0	0
	あまりお寺詣りをしない	22	45	34	7	1	0	0	109	0	0
	お寺詣りをしない	15	30	31	27	2	0	0	0	105	0
	無回答	2	0	0	0	7	0	0	0	0	9

展開したインジケータ行列との対応を示す図

$A \Rightarrow$

	1.守っている	2.まあ守っている	3.あまり守っていない	4.守っていない	5.無回答	1.お寺詣りをよくする	2.たまにお寺詣りをする	3.あまりお寺詣りをしない	4.お寺詣りをしない	5.無回答
1	1		1				1		1	
		1						1		
		1					1			
		1						1		
			1						1	
		1					1			

質問 I

質問 J

2項目の多重クロス表・バート表(B表)

バート表 対称行列	質問I	質問J
質問I	(質問I) × (質問I)のクロス表 つまり質問Iの周辺度数が対角 要素に入った対角行列	(質問I) × (質問J)の クロス表 前にみた「F表」 ①
質問J	(質問J) × (質問I)の クロス表 ② F表の転置行列	(質問J) × (質問J)のクロス表 つまり質問Jの周辺度数が対 角要素に入った対角行列

[バート表: $B = A^t A$]

多数項目になっても同じようにクロス表の並置でバート表が得られる. この関係を覚えておこう.

①と②の対応分析の結果は同じとなることを示唆.

多重クロス表と多変量モザイク図

- 多数項目の質的データのクロス表間の関連を観察する.
- 複数の2元クロス表の“並置”を行ったこと.
- “多重クロス表”を作り“個々の2項目のクロス表”からカイ二乗統計量を求めること.
- 再び環境意識調査データを例とし, ここで“複数の質問項目(多変量)”を用いてみる. 具体的には「3項目」とする.
- つまり3通りの(2元の)クロス表の組合せがある.
- これは, 多元クロス表, ここでは3元クロス表を扱うこととは異なる. あくまでも“2元データ表”である.

多変量モザイク図, 多重クロス表(バート表)

再び環境意識調査データを例とする

- Q1:あなたは、いま住んでいるまちが気に入っていますか（選択肢数4）.
- Q2:住んでいる地区は、都市としては、緑（みどり）が多いと感じますか（選択肢数5）.
- Q3:近くの緑地や公園に、どのくらい出かけますか（選択肢数5）.
- ここで「3元クロス表」ではなく「3項目（3重）の多重クロス表」であることに注意.
- ここでもピアソンのカイ二乗統計量が重要な役割を果たす.

（注）ここでは分析時に「無回答」を除いた.

多重クロス表(バート表)の観察(B表)

質問項目	質問項目 カテゴリ	まちが気に入っていますか。				緑(みどり)が多いと感じますか。					近くの緑地や公園に、どのくらい出かけますか。				
		1.大変気に入っている	2.まあ気に入っている	3.あまり気に入っていない	4.気に入っていない	1.かなり多い	2.多いほう	3.ふつう	4.少ないほう	5.少ない	1.毎日のように行く	2.週に1~2回ぐらい	3.月に1~2回ぐらい	4.年に1~2回ぐらい	5.ほとんど出かけない
まちが気に入っていますか。	1.大変気に入っている	525	0	0	0	166	239	87	26	7	39	78	167	114	126
	2.まあ気に入っている	0	1238	0	0	131	598	324	146	39	63	131	393	267	371
	3.あまり気に入っていない	0	0	172	0	6	40	55	51	20	4	20	50	23	75
	4.気に入っていない	0	0	0	15	2	2	0	5	6	0	2	4	4	5
緑(みどり)が多いと感じますか。	1.かなり多い	168	131	6	2	305	0	0	1	0	28	57	96	49	75
	2.多いほう	239	607	40	2	0	880	3	1	2	55	98	325	197	201
	3.ふつう	86	324	59	0	0	0	466	1	0	20	50	121	106	167
	4.少ないほう	26	146	51	6	0	0	0	228	0	7	17	56	42	106
近くの緑地や公園に、どのくらい出かけますか。	5.少ない	7	36	20	6	0	0	0	0	69	1	9	16	14	28
	1.毎日のように行く	39	64	4	0	27	55	18	6	1	106	0	0	0	0
	2.週に1~2回ぐらい	78	134	21	2	57	98	54	17	9	0	231	0	0	0
	3.月に1~2回ぐらい	169	394	51	4	96	325	122	59	16	0	0	614	0	0
	4.年に1~2回ぐらい	114	269	23	4	49	197	106	42	16	0	0	0	408	0
	5.ほとんど出かけない	127	375	77	6	75	201	167	106	31	5	0	0	0	577



Q1の 周辺分布	Q1×Q2の クロス表	Q1×Q3の クロス表
Q2×Q1の クロス表	Q2の 周辺分布	Q2×Q3の クロス表
Q3×Q1の クロス表	Q3×Q2の クロス表	Q3の 周辺分布

- 2元クロス表の並置に注意.
- 3元ではないこと.
- 対称行列であること.
- JMPのMCAを用いた.

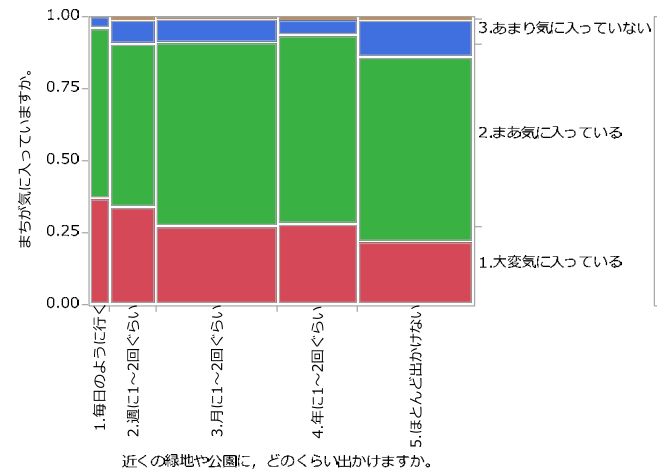
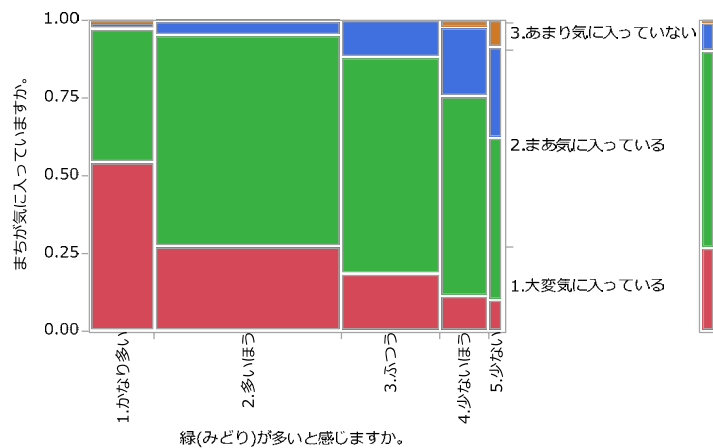
(つづき)

- 3つの質問からなる3重の“多重クロス表”となる.
- 対角ブロックが各項目の“周辺度数”(周辺分布), 非対角ブロックが2項目のクロス表からなる“対称行列”.
- これも, 2元データ表であること.
- “多元クロス表”とは異なること. ここでは3元クロス表と3重の多重クロス表は異なること.
- このデータ表の分析手法が“多重対応分析法”(MCA)である.
- 4項目以上になっても, 上下にクロス表の組合せを並置すると多重クロス表が得られる.

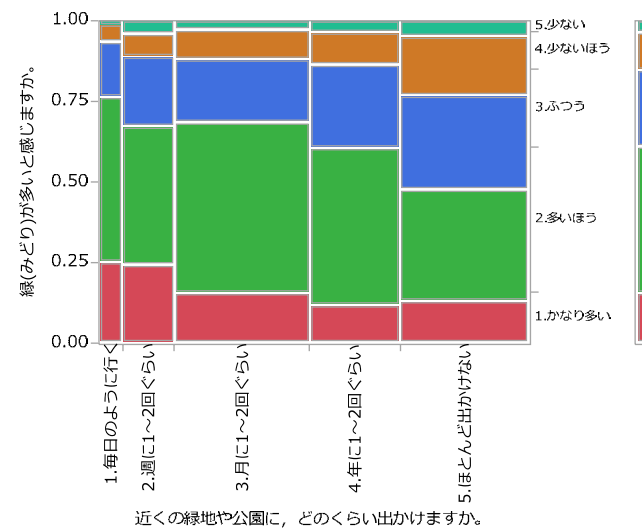
多重クロス表と多元クロス表, 対称行列

多重対応分析(MCA: multiple correspondence analysis)

モザイク図で観察(クロス表の視覚化)



Q1 × Q2の モザイク図	Q1 × Q3の モザイク図
—	Q2 × Q3の モザイク図



- 各図は上のように対応
- どれが相関が高い？

各クロス表のピアソンのカイ二乗統計量

変数名	上段: χ^2 統計量 下段: その自由度	Q1	Q2	Q3
Q1:いま住んでいるまちが気に入っていますか。	カイ二乗	–	340.309**	36.423**
	自由度	–	12	12
	測定値数	–	n=1,946	n=1,940
Q2:住んでいる地区は、都市としては、緑(みどり)が多いと感じますか。	カイ二乗	340.309**	–	102.384**
	自由度	12	–	16
	測定値数	n=1,946	–	n=1,951
Q3:近くの緑地や公園に、どのくらい出かけますか。	カイ二乗	36.423**	102.384**	–
	自由度	12	16	–
	測定値数	n=1,940	n=1,951	–

- 対角部はその質問自身なのでここでは空白とした。
- 項目の組合せにより有効データ数(n)は異なる。

平方関係係数／全慣性としてみる

変数名	上段: χ^2 統計量 下段: その自由度	Q1	Q2	Q3
Q1: いま住んでいるまちが気に入っていますか。	カイ二乗	–	0.175	0.019
	自由度	–	12	12
	測定値数	–	n=1,946	n=1,940
Q2: 住んでいる地区は、都市としては、緑(みどり)が多いと感じますか。	カイ二乗	340.309**	–	0.052
	自由度	12	–	16
	測定値数	n=1,946	–	n=1,951
Q3: 近くの緑地や公園に、どのくらい出かけますか。	カイ二乗	36.423**	102.384**	–
	自由度	12	16	–
	測定値数	n=1,940	n=1,951	–

- 対角部はその質問自身なのでここでは空白とした。
- 項目の組合せにより有効データ数(n)は異なる。

(つづき)

- 各クロス表からえたカイ二乗統計量の行列.
- 各クロス表の有効データ数と自由度が異なることに注意.
- 正確にはそのままは比べられない(参考情報となる).
- 「独立モデル」を仮定する検定結果はいづれも高度に有意となった(「**」が付いている意味;有意水準1%で有意). 当然だろう.
- 値が大きいほど(独立モデルからの乖離が大きい)関連がありそう, となる.
- これはモザイク図をみても“ある程度は類推”できる.
- ファイ係数あるいは全慣性で比べるとよい.

カイ二乗統計量の関連行列の解釈

- 3つのいずれのクロス表も, 組み合わせた2つの質問項目に“関連がありそう”となる. (いずれも「**」で有意).
- 3通りのクロス表には, その“強度に違い”がある.
「 $Q1 \times Q2$ 」(340.309) > 「 $Q2 \times Q3$ 」(102.384) > 「 $Q1 \times Q3$ 」(36.423) の順で関連度が弱まるようだ.
- 実は「 $Q1 \times Q3$ 」の値が小さい理由がある.
- 人口統計学的変数, とくに性別と年齢が関連している.

(つづき)

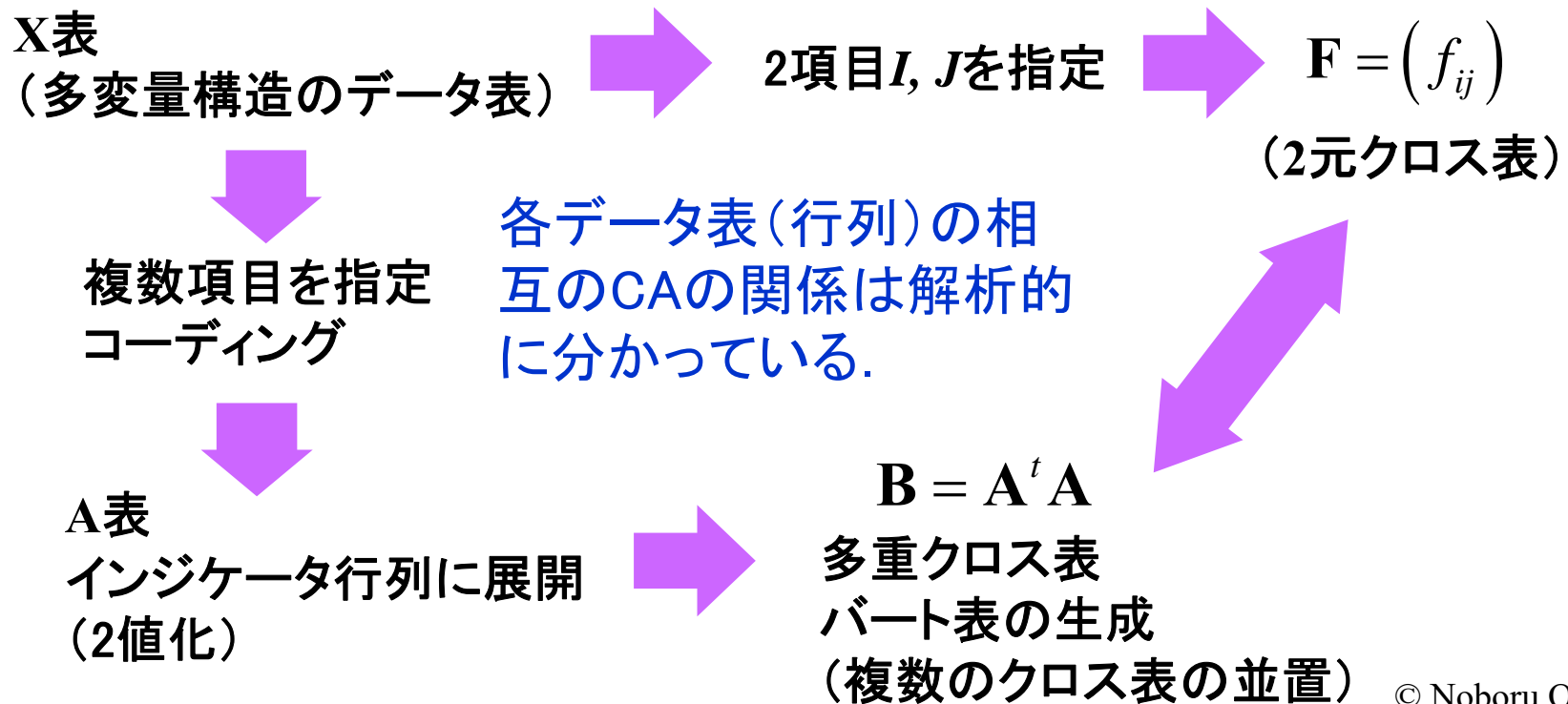
- “独立モデル”からの乖離度を測っているから、**いずれも有意はいわば当たり前**である。
- 関連(相関)があることを期待して質問文を作っているだろう。これだけで十分か？
- 標本の大きさが大きくなると、ほぼまちがいに“**有意となる**”傾向がある(有意性検定の特徴)。
- もう一步踏み込んで、クロス表の2つの項目間には具体的に“**どの程度の関連性あるいは相関があるのか**”を計量的に知りたい。

(つづき)

- これへの1つの解答が“対応分析法”あるいは“数量化Ⅲ類”を使うことで得られる.
- 2元クロス表へ“対応分析法”の適用, 多重クロス表であれば“多重対応分析法”を適用すること.
- 繰り返すが, クロス表に限らず, 条件を満たす類似の構造を持つ“2元データ表”であればよいこと.

要約: 2元データ表の相互の関連 (重要)

- テキスト「第 I 部」の最後に「付表」として整理してある情報の一部.
- 以上で述べたデータ表の相互の関係を模式図にする.
- 詳しいことは別に述べる. 「第 II 部」に説明がある.



例1: ある報告資料から

- 「旅行年報2015版」^(†)に、観光ほかに関連した多数の要約情報(2元データ表)がある.
- 「日本人の海外旅行」の項にある「旅行先別の最も楽しみにしていたこと」の要約表を取り上げて見よう(次ページに挙げた).
- ここでの課題は、この表からどのような情報抽出が可能か、「旅行先」の各地域と「楽しみ」にある各項目の間にどのような“関係がみられる”のだろうか、を知ること.
- これに対応分析法を適用して得られた結果の一部を「同時布置図」で観察する.

(†) (公益財団法人) 日本交通公社が毎年発行する報告書

旅行先別の最も楽しみにしていたこと

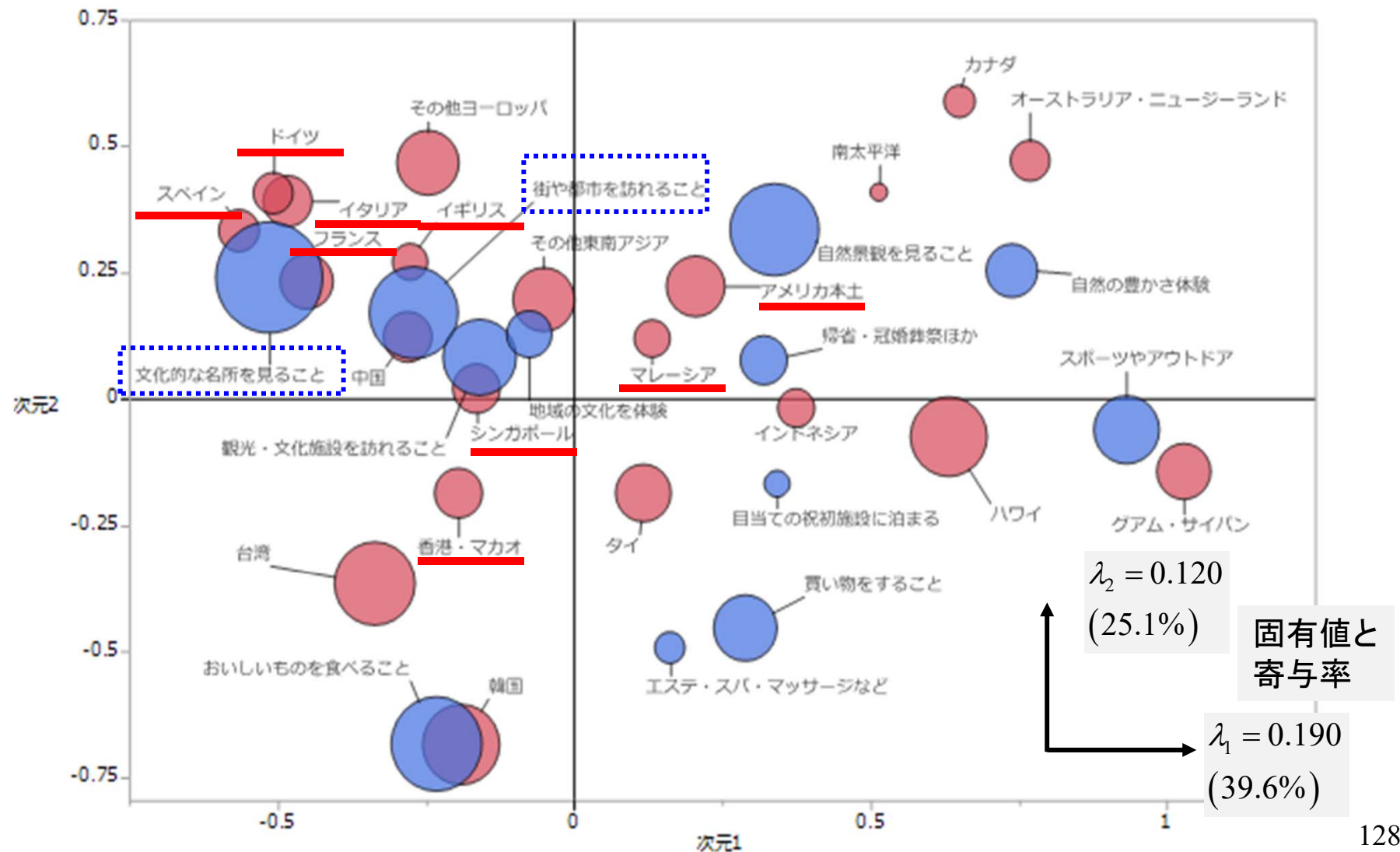
旅行先	1.文化的な 名所を見る こと	2.おいしいも のを食べる こと	3.自然景観 を見ること	4.街や都市 を訪れること	5.観光・文化 施設を訪れ ること	6.スポーツや アウトドア	7.買い物を すること	8.自然の豊 かさ体験	9.帰省・冠婚 葬祭ほか	10.地域の文 化を体験	11.エステ・ス パ・マッサージ など	12.地域の祭 りやイベント	13.目当ての 祝初施設に 泊まる	14.その他
全体	18.6	14.1	13.7	12.3	8.4	7.5	6.8	4.1	3.9	3.8	2.9	1.6	1.2	1.1
韓国	13.5	30.9	4.7	9.3	6.7	2.2	14.7	1.1	2.6	2.4	3	2.8	0.7	5.4
中国	28.8	13	10.6	13.4	6.5	2.7	3.1	2.1	6.8	5.8	0.7	1.4	1.4	3.8
台湾	20	31.8	10	12.5	9	2	4.1	1.3	2.3	2.8	0.9	0.8	0.5	2
香港・マカオ	14.9	18.8	5.7	17.4	17.7	5.3	5.7	2.1	1.8	2.5	2.1	1.1	1.8	3.2
シンガポール	12.3	10.8	7.9	20.6	20.9	2.9	5.4	1.8	4	3.6	1.4	2.2	4	2.2
インドネシア	17.4	6.4	8.7	4.7	7.6	16.9	1.7	7	5.8	5.2	11	0.6	2.3	4.7
マレーシア	11.6	10.4	14	17.7	6.7	6.1	2.4	7.9	6.1	5.5	3	0.6	1.8	6.1
タイ	20.2	16.8	6.2	10.9	2.3	15	5.2	4.4	2.3	4.9	3.9	1	1.6	5.4
その他東南アジア	31	8	14.6	10.1	2.4	9.5	3.4	2.8	4.1	5.4	3.2	0.2	1.5	3.9
オーストラリア・ニュージーランド	4.2	2.6	37.5	6.3	7.3	13.5	1.6	14.1	3.6	2.6	1	1	0.5	4.2
南太平洋	10.5	2.6	23.7	7.9	10.5	13.2	2.6	13.2	2.6	7.9	0	2.6	0	2.6
ハワイ	2.9	10.8	21	7.5	6.2	12.7	14.2	10.9	5	1.7	0.8	0.8	2.1	3.3
グアム・サイパン	2.3	7.1	12.8	4	6	35	12.8	8	4.6	2.3	0.9	0.6	1.7	2
アメリカ本土	4.2	5.1	18.5	18.3	16.2	4.2	6.9	5.1	10.4	3.7	0	2.3	0.7	4.4
カナダ	5.6	1.6	39.2	8.8	4	6.4	1.6	14.4	6.4	3.2	0	1.6	0	7.2
フランス	34.3	7.9	8.5	17.9	14.7	1.2	6.2	1.8	1.8	2.6	0.3	0.6	0	2.3
イギリス	25.3	3.9	8.4	22.1	13.6	1.9	8.4	3.9	3.2	3.9	0	0.6	0	4.5
スペイン	45.7	8	6.5	13.6	11.6	3	1.5	2	1	4.5	0	0.5	0	2
イタリア	41.5	5.9	12.1	20.1	8	2.4	3.1	2.1	0.3	2.1	0	0	0.3	2.1
ドイツ	34.5	3.4	6.4	23.6	8.9	0	3	3	3	3.9	0	4.4	1	4.9
その他ヨーロッパ	29.7	4.4	21.6	17	9.5	2.1	2.1	3.1	1.2	3.7	0	0.8	0	4.8
その他	29.4	2	26.4	12.2	3.3	7.6	2.3	4.6	1	4.3	0.7	1	0	5.3

- この表は割合、とくに行比率（行和が100）となっている。
- 列の側の比率は考えなくてもよいのだろうか。
- 報告書のコメントによる“**解釈**”は？

旅行先	文化的な名 所を見るこ と	おいしいも のを食べる こと	自然景観を 見ること	街や都市を 訪れること	観光・文化 施設を訪れ ること	スポーツや アウトドア	買い物をす ること	自然の豊か さ体験	帰省・冠婚 葬祭ほか	地域の文化 を体験	エステ・ス パ・マッサー ジなど	地域の祭り やイベント	目当ての祝 初施設に泊 まる	その他	小計
韓国	103	235	36	71	51	17	112	8	20	18	23	0	5	41	740
中国	84	38	31	39	19	8	9	6	20	17	2	0	4	11	288
台湾	158	251	79	99	71	16	32	10	18	22	7	0	4	16	783
香港・マカオ	42	53	16	49	50	15	16	6	5	7	6	0	5	9	279
シンガポール	34	30	22	57	58	8	15	5	11	10	4	0	11	6	271
インドネシア	30	11	15	8	13	29	3	12	10	9	19	0	4	8	171
マレーシア	19	17	23	29	11	10	4	13	10	9	5	0	3	10	163
タイ	78	65	24	42	9	58	20	17	9	19	15	0	6	21	384
その他東南アジア	144	37	68	47	11	44	16	13	19	25	15	0	7	18	465
オーストラリア・ニュージーランド	8	5	72	12	14	26	3	27	7	5	2	0	1	8	190
南太平洋	4	1	9	3	4	5	1	5	1	3	0	0	0	1	37
ハワイ	21	78	152	54	45	92	103	79	36	12	6	0	15	24	717
グアム・サイパン	8	25	45	14	21	123	45	28	16	8	3	0	6	7	349
アメリカ本土	18	22	80	79	70	18	30	22	45	16	0	0	3	19	422
カナダ	7	2	49	11	5	8	2	18	8	4	0	0	0	9	123
フランス	117	27	29	61	50	4	21	6	6	9	1	0	0	8	339
イギリス	39	6	13	34	21	3	13	6	5	6	0	0	0	7	153
スペイン	91	16	13	27	23	6	3	4	2	9	0	0	0	4	198
イタリア	120	17	35	58	23	7	9	6	1	6	0	0	1	6	289
ドイツ	70	7	13	48	18	0	6	6	6	8	0	0	2	10	194
その他ヨーロッパ	143	21	104	82	46	10	10	15	6	18	0	0	0	23	478
その他	89	6	80	37	10	23	7	14	3	13	2	0	0	16	300

- ここでは、各セル内の数値は、回答頻度数となっている。
- 2つの表を見て、数値の傾向が即座に読める人はかなりニューメラシーの高い人だろう。
- 対応分析では、この頻度データ表の行・列から同時的に眺めた割合(プロファイルという)を観察するので、印象が違ってくる。

対応分析で得た「同時布置図」の観察



どう観察するのか？

- ここで、「旅行先」(行)と「楽しみ」(列)の各要素の関係は、ある“傾向がある”ことが見える。
- 具体的には“どのように”読むのか(解釈するのか)。
- なぜ、ここでは(2次元の)平面内の情報として表示しているのか(なぜ、そうできたのか)。
- これで十分なのか。情報はこれで尽くされているのか。
- 布置図だけではみえない情報があるのか、それはどう調べるのか。
- データ表の情報を“視覚化”するだけでなく、さらに踏み込んで洞察するにはどうするか。



例2: メディア接触行動の分析例を比較

- 「情報に関する調査」の課題で、調査を行った.
- インターネットやソーシャル・メディアが話題となり出した90年代に、始めの調査を行った(訪問留置, 自記式).
- 2011年に、上の調査で用いた質問項目の中から、いまでも使えそうな類似項目をいくつか取り上げ、ウェブ・パネルを対象に調査を行った(ウェブ調査).
- 2つの調査から得られたデータのうち、メディア接触に関するグリッド形式の質問項目(情報接触と情報源評価)について、比較する.
- 対応分析の適用を意識して作成した質問形式.

その1:「情報に関するアンケート」の概要

- 調査対象者:18才～69才の男女(個人)
- 対象地域:東京50km圏内の市町村区
- 標本抽出:住民基本台帳(人口)による2段無作為抽出
- 調査方式:調査員が調査票配布・回収(訪問留置・回収)
- 調査期間:1997年12月6日～12月22日
- 計画標本の大きさ:1,600(人)
- 回収標本の大きさ:1,025(人)(回収率:64.1%)
- 調査機関:(株)マーケティング・サービス
- 委託社(者):A新聞社

質問項目の一部(グリッド形式)

問17. 下の表の左側に、情報を得たり交換したりできるいくつかの情報手段があげられています。

それぞれの情報手段について、あなたはどのようなイメージや印象をお持ちですか。a～sのそれぞれについて、表の上側にある選択肢の中に、あなたのイメージや印象にあてはまるものがありましたら、そのすべてに○印をつけて下さい。そのような情報手段を利用したことがない方も、持っておられるイメージでお答え下さい。
(a～sまで、それぞれ○印はいくつでも)

情報手段	イメ・ジ																
	1 おもしろい	2 親しみやすい	3 生活に欠かせない	4 役に立つ	5 身近な	6 好きな	7 嫌いな	8 かたくなるしい	9 やわらかい	10 不健全な	11 古くさい	12 将来性のある	13 特殊な	14 わかりやすい	15 くだらない	16 信頼できる	17 あてはまるものはない ひとつもない
a. テレビ(地上波)	→1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		・16
b. ケブルテレビ・衛星放送	→1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		・16
c. ラジオ	→1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		・16
d. 新聞(一般紙)	→1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		・16
e. 新聞(スポーツ紙)	→1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		・16
f. 新聞(夕刊専門紙)	→1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		・16
g. 新聞(業界専門紙)	→1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		・16

情報接触とその評価の回答集計(1996年調査)

項目	おもしろい	親しみやすい	生活に欠か せない	役に立つ	身近な	好きな	嫌いな	かたくなるしい	やわらかい	不健全な
a.テレビ(地上波)	684	432	601	504	515	286	29	13	93	38
b.ケーブルテレビ・衛星放送	377	128	70	222	88	130	8	29	25	3
c.ラジオ	338	374	179	361	337	142	17	13	76	6
d.新聞(一般紙)	128	181	573	624	380	97	9	139	14	2
e.新聞(スポーツ紙)	357	218	60	108	124	120	43	6	174	85
f.新聞(夕方専門紙)	130	112	135	236	133	34	21	42	89	43
g.新聞(業界専門紙)	46	26	63	361	27	21	19	257	9	5
h.書籍(漫画以外)	366	206	117	363	165	278	19	100	27	4
i.一般の雑誌・週刊誌(漫画以外)	417	321	72	214	227	140	27	6	176	44
j.各分野専門の情報誌	141	58	53	500	64	82	13	137	15	3
k.漫画雑誌・コミック	503	264	30	25	101	135	54	6	177	53
l.パンフレット・カタログ・ダイレクトメール	139	113	74	349	152	66	74	15	35	14
m.新聞の折り込みチラシ	146	155	205	431	276	74	29	2	46	2
n.映画・ビデオ	709	349	84	91	183	336	10	6	61	10
o.CD・MD	315	328	142	77	207	316	9	6	72	5
p.テープ・LP	257	274	92	73	187	260	7	9	50	5
q.電話・ファックス情報サービス	32	24	51	336	53	12	14	24	10	12
r.インターネット・パソコン通信	135	45	58	306	47	44	19	48	15	11
s.CD-ROMや電子ブック	75	25	28	240	21	26	21	49	6	3
列和	5293	3632	2686	5419	3285	2598	444	908	1173	351

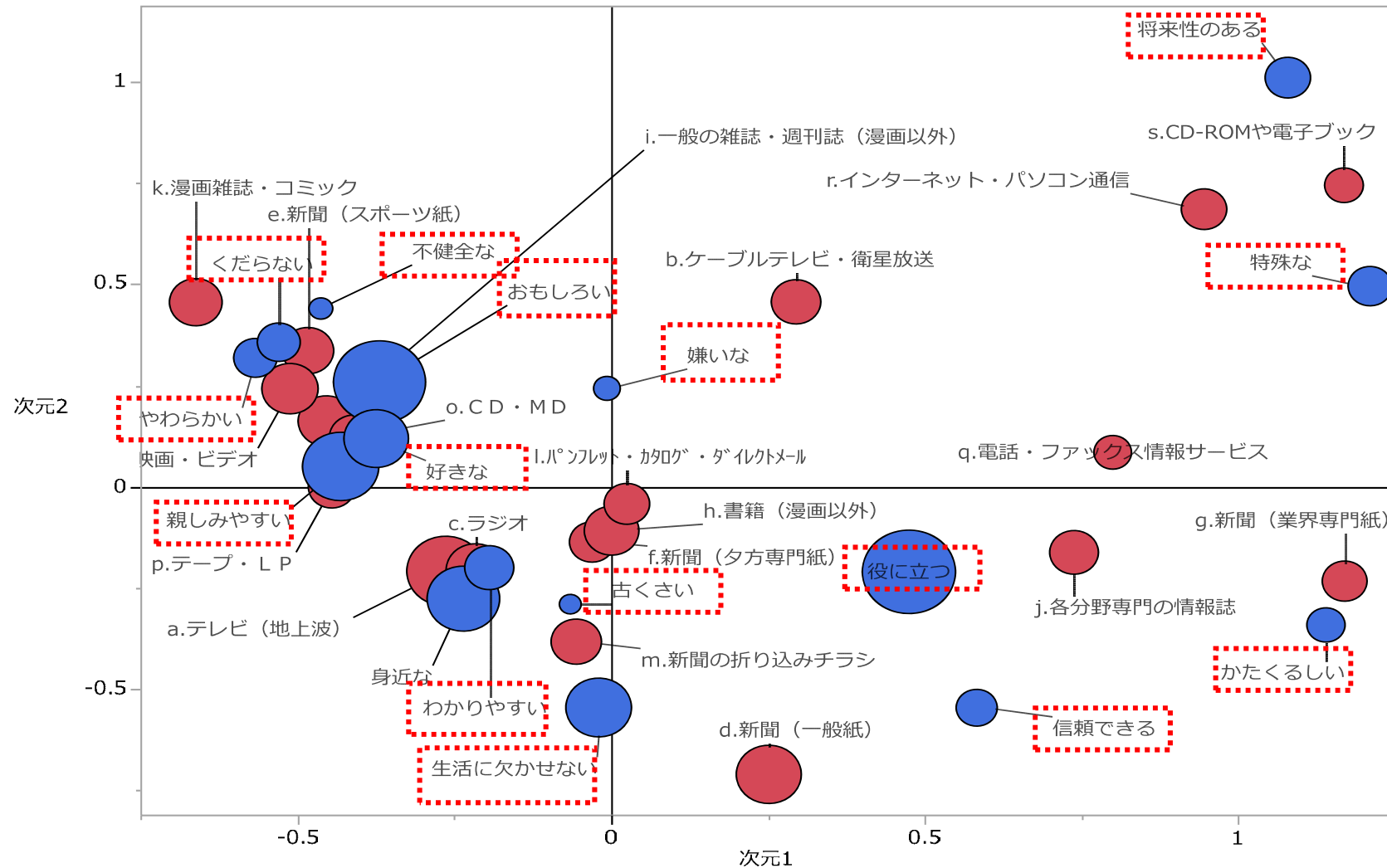
「17. あてはまるものはない」「無回答」は除外した.

(つづき)

項目	古くさい	将来性のある	特殊な	わかりやす	くだらない	信頼できる
a.テレビ(地上波)	6	53	8	285	159	80
b.ケーブルテレビ・衛星放送	2	218	108	51	15	52
c.ラジオ	49	23	10	125	33	78
d.新聞(一般紙)	24	34	8	150	11	263
e.新聞(スポーツ紙)	9	7	53	93	194	16
f.新聞(夕方専門紙)	14	11	42	63	96	60
g.新聞(業界専門紙)	23	30	222	19	15	110
h.書籍(漫画以外)	17	33	30	46	10	78
i.一般の雑誌・週刊誌(漫画以外)	6	15	17	126	158	19
j.各分野専門の情報誌	8	46	170	56	10	122
k.漫画雑誌・コミック	3	9	17	100	247	5
l.パンフレット・カタログ・ダイレクトメール	9	30	39	77	95	15
m.新聞の折り込みチラシ	18	7	13	123	35	19
n.映画・ビデオ	7	31	17	60	10	11
o.CD・MD	2	62	17	31	7	11
p.テープ・LP	81	13	15	27	5	8
q.電話・ファックス情報サービス	6	103	96	43	21	42
r.インターネット・パソコン通信	6	329	149	19	13	37
s.CD-ROMや電子ブック	3	232	167	8	10	22
列和	295	1286	1201	1503	1147	1050

この「2元データ表」(19行×16列)に対応分析を適用する。

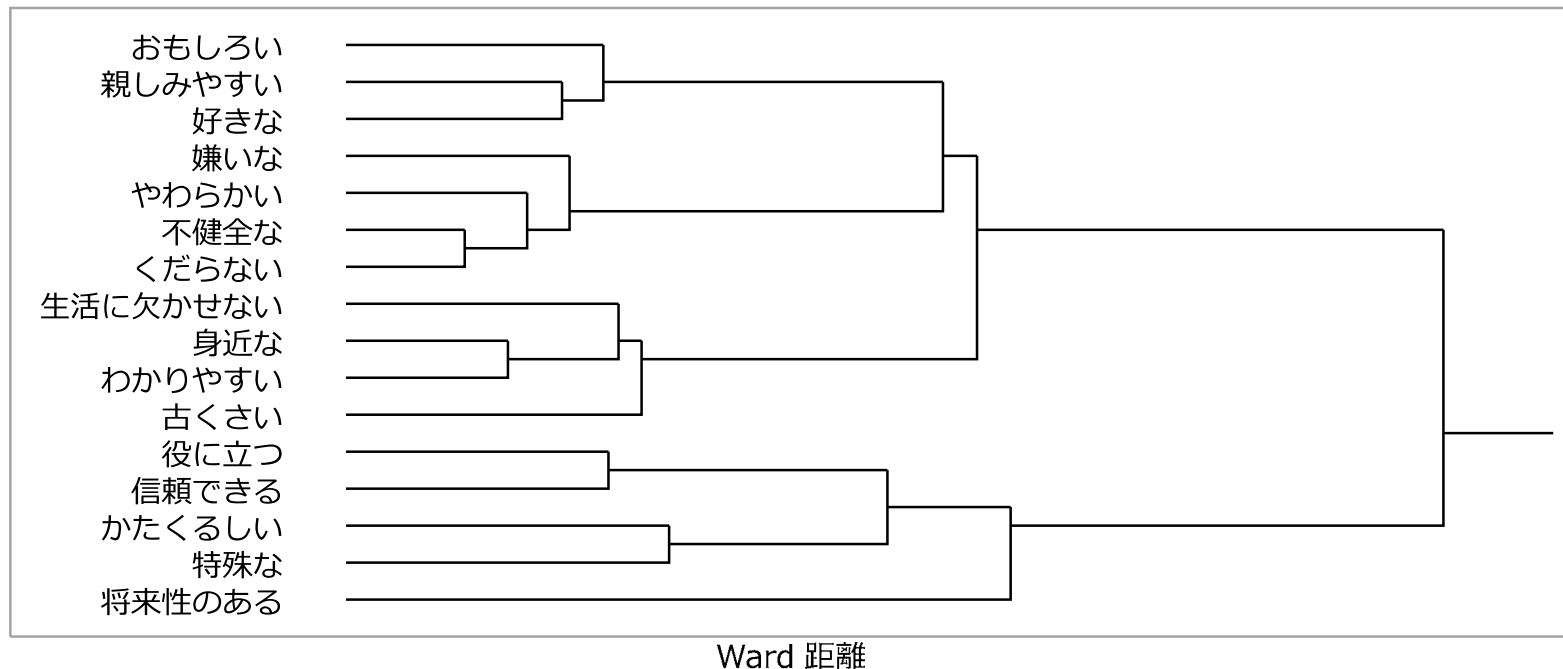
同時布置図の観察(留置・自記式の場合)



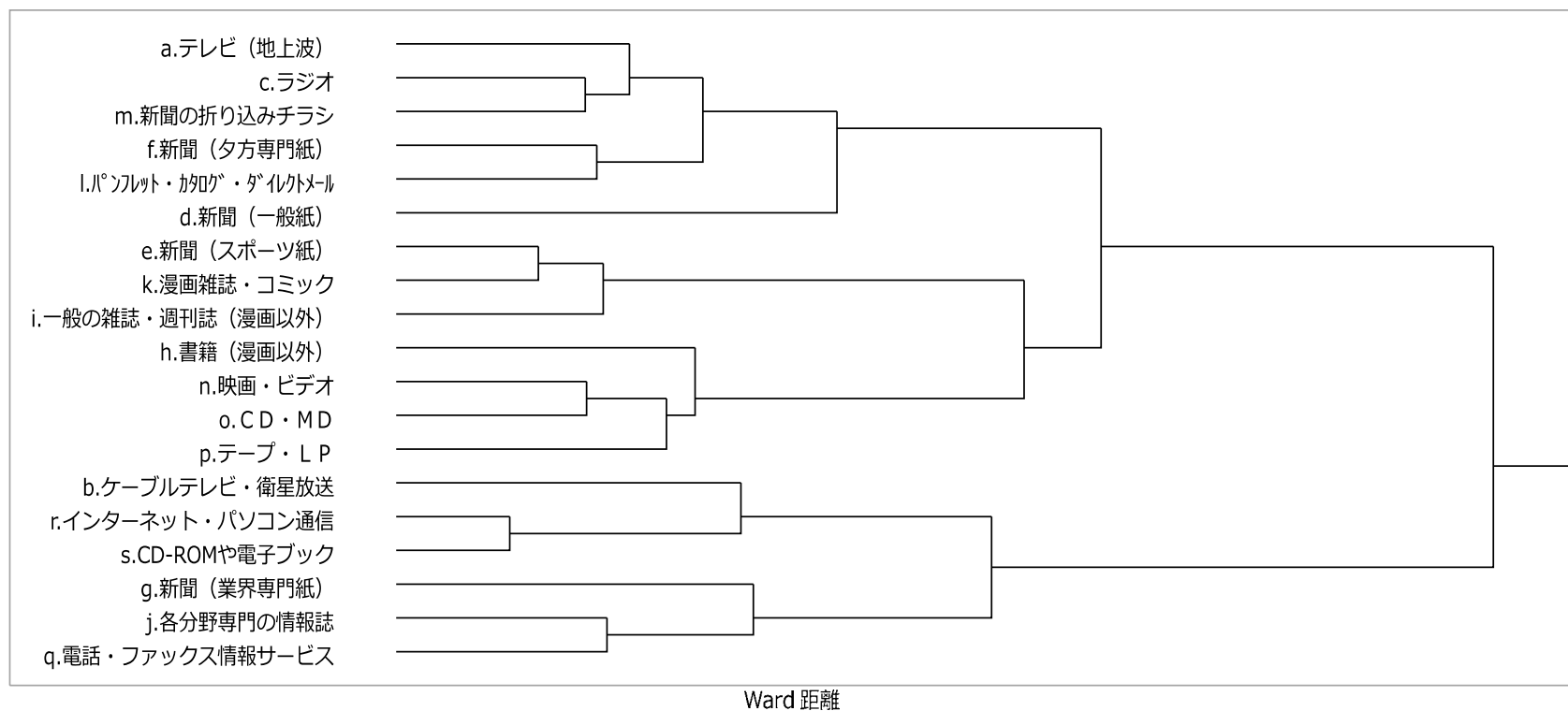
メディアと評価項目の関係がみえる(細かいことは略)

行, 列それぞれの樹形図(デンドログラム)

- 評価項目の分類を行い, 以下を得た.
- ここで, “Ward距離”(ウォード距離)とあることに注意する.
- 分類時のクラスター間距離にウォード距離を用いたと言うこと. これが対応分析とどう関連するのか.



● 接触したメディア項目の分類の結果.



その2:「情報に関する調査」の概要

- インターネット時代になって、どう状況が変わったのか.
- 2011年に、あるウェブ・パネルを用いて、ウェブ調査により類似の調査を行ってみた.
- 取り上げるメディアの環境は大きく変わっているので、同じ質問項目や選択肢を用意することは無理である.
- 接触メディアを変えて、また評価項目はなるべく共通するものを選んで、実施した.
- その結果の1つが、パブリシティ用資料にある事例データに相当する.

[調査の概要]

- 調査課題: 情報に関する調査
- 調査対象(標本抽出枠): あるウェブ・パネル(非公募型)に登録の首都40km圏に在住, 15歳以上69歳未満の男女(パネル構成の詳細情報は省く)
- 調査方式(モード): ウェブ調査
- 実施期間: 2011年9月9日(17時)～9月13日(10時)まで
- 計画標本の大きさは766(人), 有効回収標本の大きさは347(人), 参加率は45.3(%) [注: 参加率が高いとはいえない]

非公募型 ⇨ 部分的に確率標本, ということ

分析に用いる質問項目（調査票の部分）

Q17 現在私たちは、情報入手できる手段として数多くの情報源に囲まれており、それらの情報源についていろいろな意見が言われています。
さて、以下でAからDの4つのことばがあてはまる情報源にはどのようなものがあるでしょうか。
あなたが「あてはまる」と思われるものをすべてお選びください。（それぞれいくつでも）

	A 情報 が正 確	B 情報 が詳 しい	C 情報 量が多 い	D 信頼 できる
	↓	↓	↓	↓
1. テレビの番組	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. ケーブルテレビ・衛星放送の番組	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. ラジオの番組	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. 新聞の記事（電子版を含む）	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. 新聞の紙面広告（電子版を含む）	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. 書籍（漫画・コミック以外）	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. 一般の雑誌・週刊誌（漫画・コミック以外）の記事	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8. 各分野専門の情報誌の記事	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9. パンフレット・カタログ・ダイレクトメール	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10. 都・県や市・区など自治体の広報誌紙	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11. 所属する会や組織の会報・同人誌・ニュースレター	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

グリッド（マトリクス）形式
チェックボックスを利用

情報源(23選択肢)と評価項目(9選択肢)

情 報 源		評 価 項 目
1. テレビの番組	12. パソコンでみるインターネットサイト	情報が正確
2. ケーブルテレビ・衛星放送の番組	13. 携帯電話・PHS、スマートフォンでみるインターネットサイト	情報が詳しい
3. ラジオの番組	14. インターネットブログ、ブログ	情報量が多い
4. 新聞の記事(電子版を含む)	15. ツイッター(Twitter)	信頼できる
5. 新聞の紙面広告(電子版を含む)	16. 電子書籍(電子書籍端末や電子ブックリーダーで読む)	生活に欠かせない
6. 書籍(漫画・コミック以外)	17. ミクシィ(mixi)	役に立つ
7. 一般の雑誌・週刊誌(漫画・コミック以外)の記事	18. フェイスブック(Facebook)	世間の話題や流行を知る
8. 各分野専門の情報誌の記事	19. グリー(GREE)	商品を選び購入する
9. パンフレット・カタログ・ダイレクトメール	20. モバゲータウン	古くさい
10. 都・県や市・区など自治体の広報誌紙	21. YouTube	
11. 所属する会や組織の会報・同人誌・ニュースレター	22. ニコニコ動画	

$$m = 22, n = 9, K = \min\{22, 9\} - 1 = 8$$

(こういう大きさの2元データ表)

分析対象とする2元データ表

表 53 [情報源(23 選択肢)] × [評価項目(9 項目)]の 2 元データ表

質問項目	Q17_A 情報が正確	Q17_B 情報が詳しい	Q17_C 情報量が多い	Q17_D 信頼できる
全サンプル数（回答者数）	347	347	347	347
1. テレビの番組	100	114	235	90
2. ケーブルテレビ・衛星放送の番組	43	91	103	44
3. ラジオの番組	65	68	86	54
4. 新聞の記事（電子版を含む）	140	153	131	131
5. 新聞の紙面広告（電子版を含む）	36	59	90	28
6. 書籍（漫画・コミック以外）	41	98	108	39
7. 一般の雑誌・週刊誌（漫画・コミック以外）の記事	19	84	139	12
8. 各分野専門の情報誌の記事	88	163	89	86
9. パンフレット・カタログ・ダイレクトメール	31	90	90	18
10. 都・県や市・区など自治体の広報誌紙	125	70	38	132
11. 所属する会や組織の会報・同人誌・ニュースレ	45	73	50	46
12. パソコンでみるインターネットサイト	26	118	274	24
13. 携帯電話・PHS、スマートフォンでみるインターネットサイト	18	63	166	15
14. インターネットブログ、ブログ	6	59	150	5
15. ツイッター（Twitter）	6	31	143	9
16. 電子書籍（電子書籍端末や電子ブックリーダー	22	39	103	19
17. ミクシィ(mixi)	9	36	128	11
18. フェイスブック（Facebook）	12	17	115	12
19. グリー（GREE）	6	20	99	5
20. モバゲータウン	8	20	100	4
21. YouTube	16	42	135	10
22. ニコニコ動画	7	22	122	6
23. 1～22 の中にはひとつもない	54	23	11	62
99. 無回答	0	0	0	0

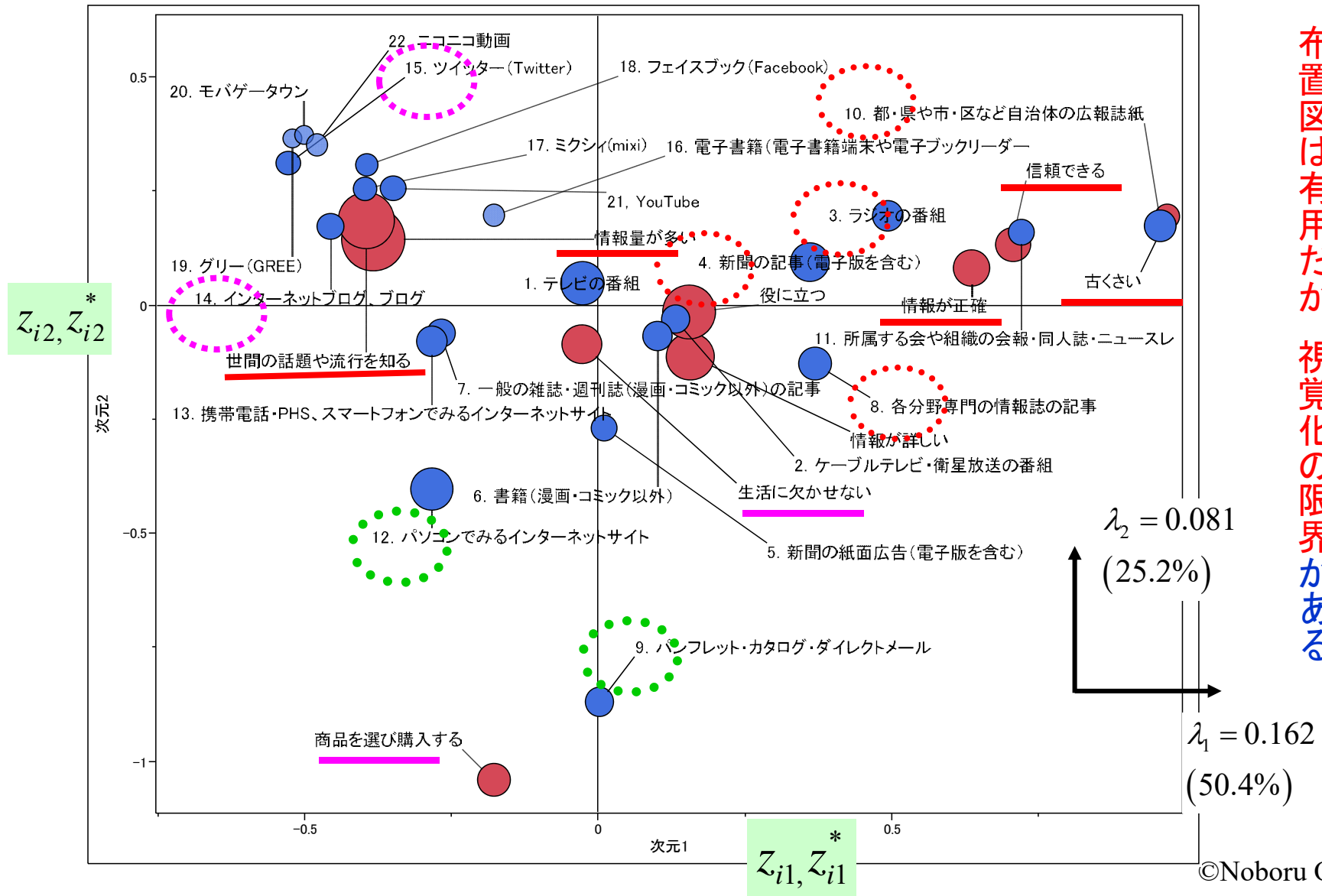
(つづき)

[情報源(23 選択肢)] × [評価項目(9 項目)]の2元データ表(つづき)

質問項目	Q18_A- 生活に欠か せない	Q18_B- 役に立つ	Q18_C- 世間の話題 や流行を知 る	Q18_D- 商品を選び 購入する	Q18_E- 古くさい
サンプル数	347	347	347	347	347
1. テレビの番組	226	166	248	51	20
2. ケーブルテレビ・衛星放送の番組	42	94	74	30	17
3. ラジオの番組	50	103	79	9	82
4. 新聞の記事(電子版を含む)	135	167	124	17	24
5. 新聞の紙面広告(電子版を含む)	29	62	74	59	16
6. 書籍(漫画・コミック以外)	57	100	71	33	16
7. 一般の雑誌・週刊誌(漫画・コミック以外)の記事	27	74	136	46	12
8. 各分野専門の情報誌の記事	24	137	63	47	10
9. パンフレット・カタログ・ダイレクトメール	12	76	59	155	29
10. 都・県や市・区など自治体の広報誌紙	41	148	25	7	79
11. 所属する会や組織の会報・同人誌・ニュースレ	11	102	25	5	67
12. パソコンでみるインターネットサイト	186	204	200	182	0
13. 携帯電話・PHS、スマートフォンでみるインターネットサイト	77	105	107	44	0
14. インターネットブログ、ブログ	34	67	135	14	1
15. ツイッター(Twitter)	21	43	117	5	1
16. 電子書籍(電子書籍端末や電子ブックリーダー	9	53	61	7	1
17. ミクシィ(mixi)	25	43	107	8	8
18. フェイスブック(Facebook)	14	35	94	7	9
19. グリー(GREE)	6	18	79	3	6
20. モバゲータウン	7	20	76	2	6
21. YouTube	27	88	120	4	1
22. ニコニコ動画	10	46	94	2	5
23. 1~22の中にはひとつもない	25	12	13	38	123
99. 無回答	0	0	0	1	0

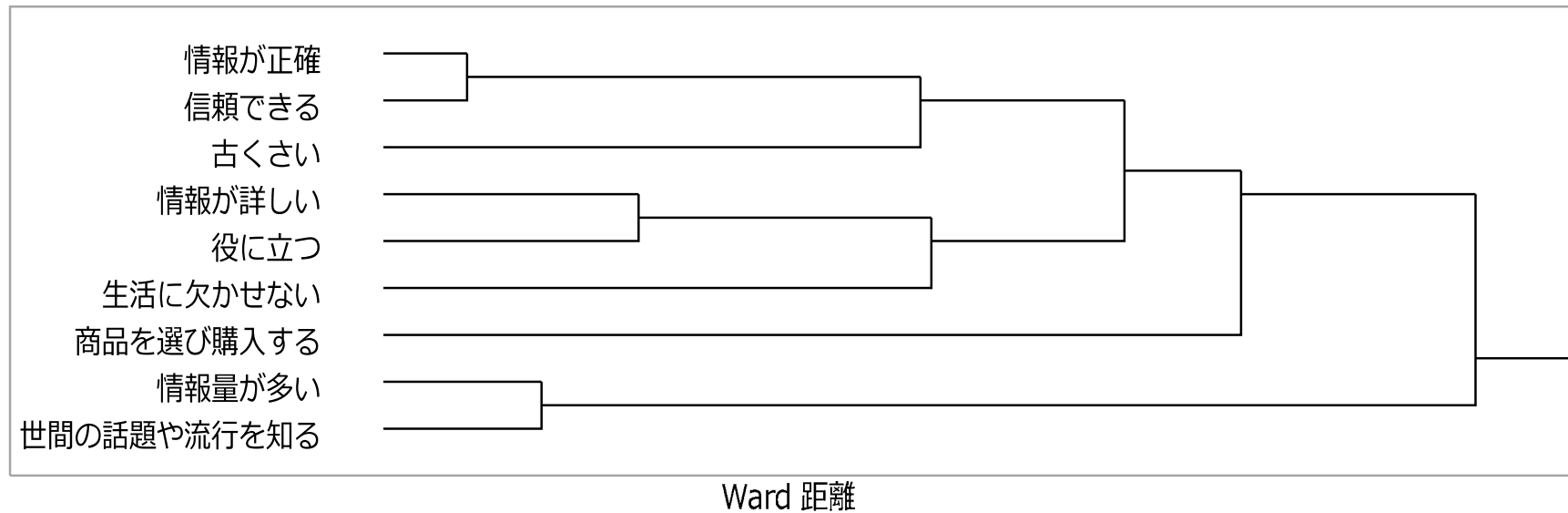
ここで「22. この中にひとつもない」「99. 無回答」は除外した。
含めると、どうなるだろうか？

同時布置図の観察(ウェブ調査の場合)



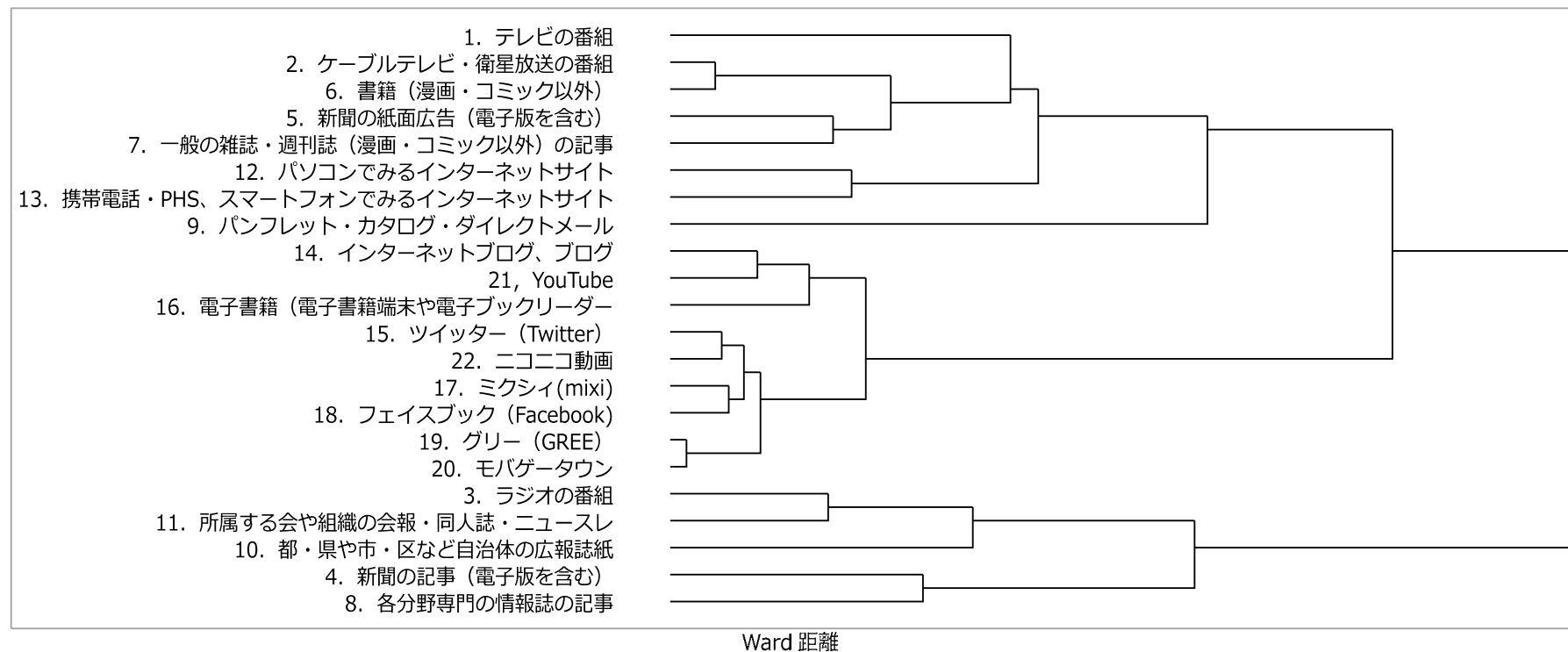
行, 列それぞれの樹形図(デンドログラム)

- ここでも, 評価項目の分類を行い, 以下を得た.



(つづき)

- 接触したメディア項目の分類の結果
- ここで、右端をカット，距離が遠い，つまり離れている。



【補足】「全国学力調査」(2014)から

【小学校調査】

教科	平均正答数	平均正答率
国語A	11.0 問／15 問	73.1%
国語B	5.6 問／10 問	55.6%
算数A	13.3 問／17 問	78.2%
算数B	7.6 問／13 問	58.4%

【中学校調査】

教科	平均正答数	平均正答率
国語A	25.5 問／32 問	79.8%
国語B	4.6 問／09 問	51.6%
数学A	24.5 問／36 問	67.9%
数学B	9.1 問／15 問	60.5%

(*) 文部科学省のサイトから得た情報を要約



「正答数」に注目

全国学力・学習状況調査(小学校) 県別 - JMP

ファイル(F) 編集(E) テーブル(T) 行(R) 列(C) 実験計画(DOE)(D) 分析(A) グラフ(G) ツール(Q) アドイン(N) 表示(V) ウィンドウ(W) ヘルプ(H)

全国学力・...

列(9/0)

県名

国語A (正答数)

国語B (正答数)

算数A (正答数)

算数B (正答数)

国語A (正答率)

国語B (正答率)

算数A (正答率)

算数B (正答率)

行

すべての行 48

選択されている行 0

除外されている行 0

表示しない行 0

ラベルのついた行 0

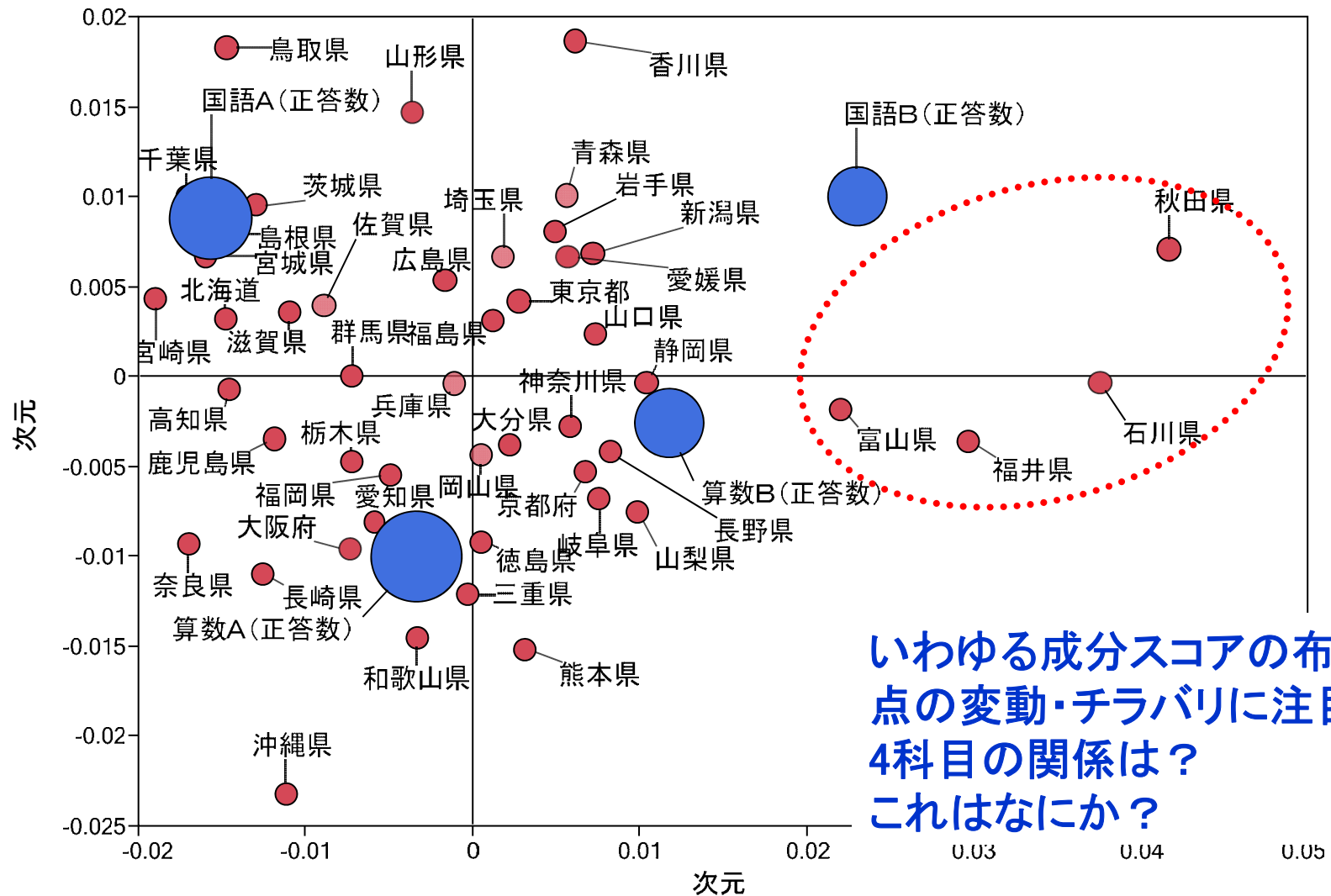
県名	国語A (正答数)	国語B (正答数)	算数A (正答数)	算数B (正答数)	国語A (正答率)	国語B (正答率)	算数A (正答率)	算数B (正答率)
1 北海道	10.8	5.3	12.9	7.2				
2 青森県	11.5	6	13.8	7.9				
3 岩手県	11.1	5.8	13.4	7.6				
4 宮城県	11.1	5.4	13.1	7.4				
5 秋田県	11.6	6.7	14.5	8.6				
6 山形県	11.2	5.7	13.2	7.5				
7 福島県								
8 茨城県								
9 栃木県	10.8	5.4	13.2	7.4				
10 群馬県	11	5.5	13.2	7.5				
11 埼玉県	10.9	5.6						
12 千葉県	11.4	5.5						
13 東京都	11.3	5.7						
14 神奈川県	10.7	5.5	13.1	7.6				
15 新潟県	11.2	5.9	13.6	7.7				
16 富山県	11.1	6	13.8	8.1				
17 石川県	11.1	6.3	14	8.3				
18 福井県	11.2	6.2	14.1	8.3				
19 山梨県	10.5	5.5	13.1	7.5				

(県 × 4科目 正答数)の2元データ表

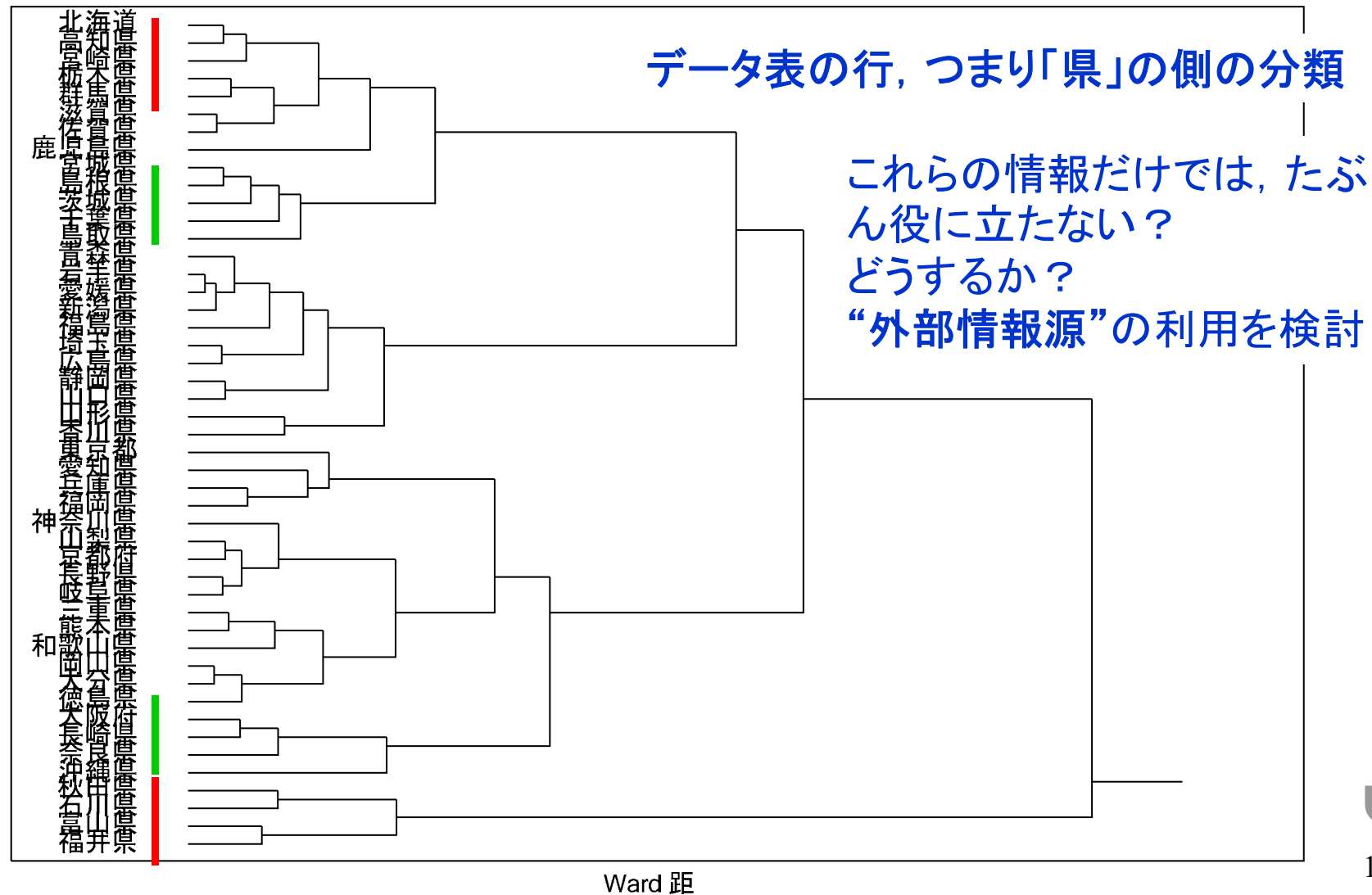
なにが“質的データ”なのか？
なにを“分析”するのか？



県名と4科目の同時布置図



分類結果の樹形図(デンドログラム)



JMRAマーケティング・リサーチ講座

質的データのマイニング のための対応分析法

【補足】

記述的統計量とは
とくに、平均、分散と標準偏差
標準化の操作

※「探索発見的データ解析」講座
資料から、一部を抜粋再編集した

大隅 昇

ohsumi@ss.iij4u.or.jp

<http://wordminer.org/>

Copyright by Noboru Ohsumi

なぜ、記述的統計量を用いるか？

- 記述的統計量とは，“データ探査”に用いる指標群のこと.
- たとえば，平均（平均値），中央値，分散・標準偏差，…
- データ探査の基本は，データの示す“分布の特徴・傾向”を知ること.
- （個々の測定値ではなく）集めた集団としての傾向を知る.
- “分布”とはなにか．量的データであれば，ドットプロット図やヒストグラムを描いて観察される“姿”のこと.
- “分布”を測る記述的統計量は，“無数”の種類がある.
- 換言すると，“測っている特徴”（≡意味）が違うということ.

データ，測定値・観測値などを混用，
意味は（ほぼ）同じ.

代表的な記述的統計量には, …

- 分布の“位置”を測る指標
 - 例: 平均, 中央値, モード(最頻値)など
- 分布の“変動”(バラツキ, チラバリ), “形”を測る指標
 - 分散・標準偏差, 四分位範囲(ヒンジ散布度)
- 分布の“歪み”, “偏り”を測る指標: 歪度
- 分布の“尖り”, “集中度”を測る指標: 尖度
- とくに, 分布の位置と変動を知ることが重要.
- 統計量とグラフィカル表現法の併用が重要.
- “目で見ると同時に“数値で確認”とする.
- 錯誤, 誤用の回避の意味がある.

ここで知っておく記述的統計量

- “量的データ” (区間尺度, 比例尺度) に用いる指標に限定する
- 原則, 四則演算が可能 (除算は比例尺度のみ).
- 次の“統計量” だけを調べる.
 - 平均または重心 (セントロイド)
 - 偏差と偏差平方和 (平方和)
 - 平方和と分散, 標準偏差
- 統計量は測定値のなにを測っているのかを確認する.

1変量の“量的”データ(測定値)

個体番号 (i)	変量 (x_i)
1	x_1
2	x_2
3	x_3
\vdots	\vdots
i	x_i
\vdots	\vdots
$n-1$	x_{n-1}
n	x_n

$$\Leftrightarrow \begin{cases} x_1, x_2, \dots, x_i, \dots, x_{n-1}, x_n \\ \text{または} \\ x_i \quad (i = 1, 2, \dots, n) \end{cases}$$

記法を覚える.

「添え字」がカウンターとなる

大きさが n の1変量データをこのように表す(測定値, 観測値)

トイ・データによる確認(1変量の量的データ)

- 何かの特性を測定したような場面を想定する.
- “大きさ”が $n=26$ の“1変量”データという.
- ここでは“量的データ”と考える.
- 測定単位があるとして, たとえばその測定単位を「グラム(g)」としよう.
- まず, “データ表”に整理し, 確認する.

1変量(単変量), データ表



(つづき)

個体番号 (i)	変量 (x_i)	個体番号 (i)	変量 (x_i)	個体番号 (i)	変量 (x_i)
1	4	11	5	21	4
2	7	12	7	22	3
3	10	13	9	23	4
4	8	14	5	24	2
5	7	15	2	25	6
6	5	16	4	26	3
7	9	17	5		
8	8	18	7		
9	4	19	6		
10	6	20	5		



「ドットプロット図」による観察

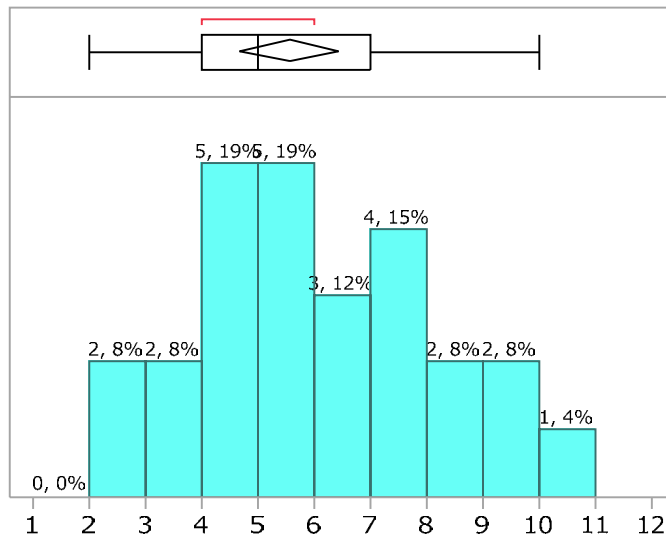


再確認:

- “ドットプロット図” (打点図) による観察, 傾向探査.
- 1 変量の量的データの観察に適している.
- ヒストグラムとは異なる.
- 度数 (頻度) が分かる.
- 測定単位とその精度を知る.
- “分布” の姿を知る (位置, 広がり・形, はずれ値など).

JMPによる出力例

- “ヒストグラム”, “箱ひげ図”ほかのグラフィカル表現
- 基本的な要約統計量(記述的統計量)の表示
- 図中にも統計量が表示されている(平均, 中央値, 四分位範囲, 最短の半分(SHORTH), 中央絶対偏差など).



分位点			要約統計量	
100.0%	最大値	10	平均	5.5769231
99.5%		10	標準偏差	2.1572775
97.5%		10	平均の標準誤差	0.4230769
90.0%		9	平均の上側95%	6.4482663
75.0%	四分位点	7	平均の下側95%	4.7055798
50.0%	中央値	5	N	26
25.0%	四分位点	4	合計	145
10.0%		2.7	分散	4.6538462
2.5%		2	歪度	0.2617318
0.5%		2	尖度	-0.602132
0.0%	最小値	2	変動係数	38.682217
			欠測値 N	0
			無修正平方和	925
			修正平方和	116.34615
			最小値	2
			最大値	10
			中央値	5
			範囲	8
			四分位範囲	3
			中央絶対偏差	1.5

とくに, 平均, 分散と標準偏差, (修正)平方和の各値に注目, 覚えておこう.

基本の記述的統計量の確認

- ここから、平均(標本平均)、分散(標本分散、不偏分散)、標準偏差という、基本的な“**積率型の統計量**”の使い方と性質を述べる.
- 統計利用の上で、もっとも日常的に使われていながら、実は正確に理解されていない部分がある.
- 平均は測定値の示す分布の位置を示す代表値の1つ.
- 分散、標準偏差は、データの分布の変動(バラツキ, チラバリ)や形を測る指標.
- これらとその元になる統計量(平方和、偏差など)の性質を圧縮して説明する.

積率型の統計量

- 積率(標本積率)とは, 平均を一般化したことという程度の意味に考えておく.
- つまり, “平均”をより広い意味として考えること.
- “平均”とは, “平らに均す”こと.
- 積率をモーメント(動的慣性)ともいう.

$\frac{1}{n} \sum_{i=1}^n (\cdot)^k$ の平均を作る形で書ける統計量を“積率”型という
この $(\cdot)^k$ の中に“何が入るか”, それが意味するものを知る
どれもが“平均”ということ.

準備：積率型の統計量の“例”

$$\alpha_k = \frac{1}{n} \sum_{i=1}^n x_i^k \quad \left(\begin{array}{l} \text{原点のまわりの} k \text{ 次の積率} \\ \text{測定値} x_i \text{ の} k \text{ 乗の平均} \end{array} \right)$$

$$\beta_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k \quad \left(\begin{array}{l} \text{平均} \bar{x} \text{ のまわりの} k \text{ 次の積率} \\ \text{平均} \bar{x} \text{ のまわりの偏差} (x_i - \bar{x}) \text{ の} k \text{ 乗の平均} \end{array} \right)$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (= \alpha_1) \quad (\text{平均} = \text{原点のまわりの} 1 \text{ 次の積率})$$

$$s_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (= \beta_2) \quad (\text{分散} = \text{平均値のまわりの} 2 \text{ 次の積率})$$

統計量(標本統計量)とは何か(の説明)が重要だが略す.
ここでは, 測定値(x_i)から作った関数, 位に考えておく.

さらに, ...

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \quad (= \beta_1) \left(\begin{array}{l} \text{偏差の平均} \\ \text{平均}\bar{x}\text{のまわりの1次の積率} \end{array} \right)$$

$$\nu_k = \frac{1}{n} \sum_{i=1}^n (x_i - a)^k \left(\begin{array}{l} \text{ある定数}a\text{のまわりの}k\text{次の積率} \\ \text{偏差}(x_i - a)\text{の}k\text{乗の平均} \end{array} \right)$$

$$\textcircled{1} \text{ここで, } a = 0 \text{ のとき} \Rightarrow \alpha_k = \frac{1}{n} \sum_{i=1}^n x_i^k$$

$$\textcircled{2} \text{ここで, } a = \bar{x} \text{ のとき} \Rightarrow \beta_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$$

- 共通の表記法 (一般化) で記述していることに注意する.
- (経験) 積率を使って統計量を考える方法を (標本) 積率法 (モーメント法) という.
- 統計量の作り方は この方式だけではない (無数にある).

以下の統計量だけを取り上げる

- 実用上は限られた意味のわかる(解釈が容易な)統計量を用いている. とりあえずそれで十分.
- ①(狭い意味の)平均あるいは重心(セントロイド)
- ②偏差(平均のまわりの偏差)
- ③ 偏差平方和あるは平方和
- ④分散(標本分散), 不偏分散(不偏な標本分散)
- ⑤標準偏差(分散の正の平方根)
- それぞれの統計量が分布の“何を説明”するのかを知っておくことが必要(目標).
- トイ・データを用意し, 具体的に説明する.

1変量の“量的データ”の記法

個体番号 (i)	変量 (x_i)
1	x_1
2	x_2
3	x_3
\vdots	\vdots
i	x_i
\vdots	\vdots
$n-1$	x_{n-1}
n	x_n

大きさが n の1変量データ(測定値, 観測値)をこのように表す.

$$\Leftrightarrow \begin{cases} x_1, x_2, \dots, x_i, \dots, x_{n-1}, x_n \\ \text{または} \\ x_i \quad (i = 1, 2, \dots, n) \end{cases}$$

記法を覚えよう.

- 「添え字」をカウンターとする
- 一般に英(小)文字で測定値
- 小文字と大文字は使い分ける
(意味が異なるので注意 \Leftrightarrow ガイダンス)

①「平均」

- ここでは、平均として“算術平均”（相加平均）とする.
- 大きさが n の1変量データ(ここは量的データ)の平均.
- つまり“**原点のまわりの1次の積率**”のこと.

$$\bar{x} = \frac{1}{n} \underbrace{(x_1 + x_2 + \cdots + x_i + \cdots + x_n)}_{\text{総和をとる}} = \frac{1}{n} \underbrace{\sum_{i=1}^n x_i}_{\text{この記法(演算子)に注意!!!}}$$

「エックスバー」
と読む

総和をとる

この記法(演算子)に注意!!!

$$\frac{1}{n} \sum_{i=1}^n (\cdot)^k \text{ の形}$$

Σ : 総和記号(シグマ: sum)
ギリシャ文字のSの大文字

平均(mean, average), 重心(centroid, center of gravity)

例で確認しよう

実際にこの大きさが26のデータから平均を求める.

$$\bar{x} = \frac{1}{26} (4 + 7 + \cdots + 6 + 3) = \frac{1}{26} \times 145 = \underline{5.577} \text{ (g)}$$

- 測定の単位 (g) は元のまま, 変化はない.
- 求める数値の桁数は小数点以下2桁 (有効桁数は?).
もとの測定値の桁数に対して何桁まで有効か?
- とりあえず測定値の分布の位置の目安は得られた.
- では“変動” (バラツキ, チラバリ) はどう測るのか?

②「偏差」を考える

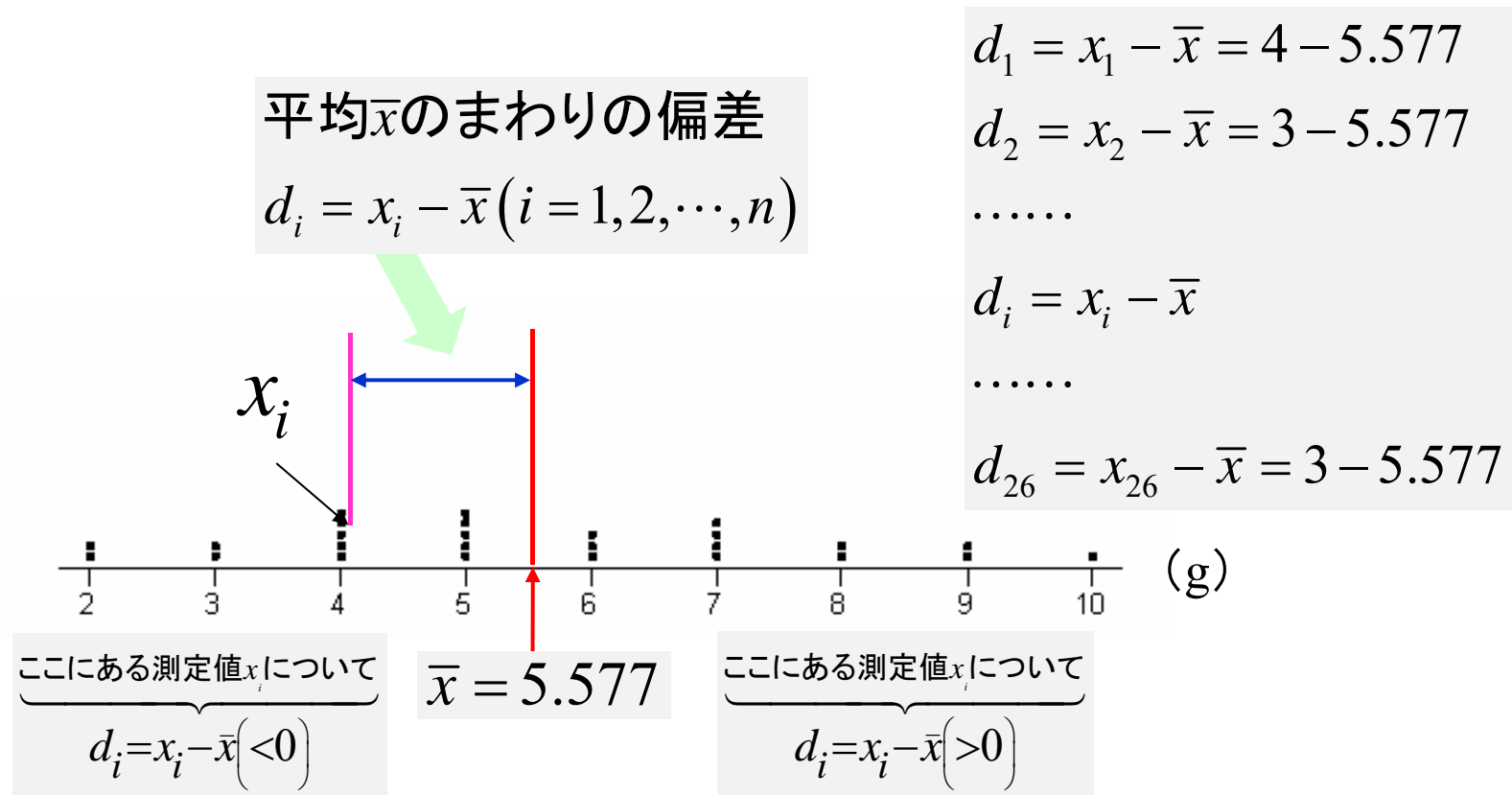
- 偏差はいろいろある. ここは「平均のまわりの“偏差”」を考える.
- 偏差とは, 測定値と平均との“ズレ”のこと.
- つまりある種の“距離”(符号を持った距離).
- 個々の測定値の“変動”(バラツキ, チラバリ)を測る指標のように思える.

平均 \bar{x} のまわりの偏差

$$d_i = x_i - \bar{x} \quad (i = 1, 2, \dots, n)$$

偏差(deviation), ずれ(discrepancy, gap)

偏差をドットプロット図の上で観察



- 平均のまわりでセンタリング(中心化)という.
- 平均で分布の位置は分かるが, 変動(バラツキ, チラバリ)は?

平均のまわりの偏差の和と平均(重要)

偏差の和:

$$\sum_{i=1}^n d_i = \sum_{i=1}^n \underbrace{(x_i - \bar{x})}_{\geq 0, \leq 0} = \sum_{i=1}^n x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$$

偏差の平均:

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = \frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \times n\bar{x} = \bar{x} - \bar{x} = 0$$

- 平均値 \bar{x} のまわりの偏差の和は「ゼロ」である.
- 平均値のまわりの偏差の平均は「ゼロ」である.
- 上の式は“平均値のまわりの1次の積率”ということ.

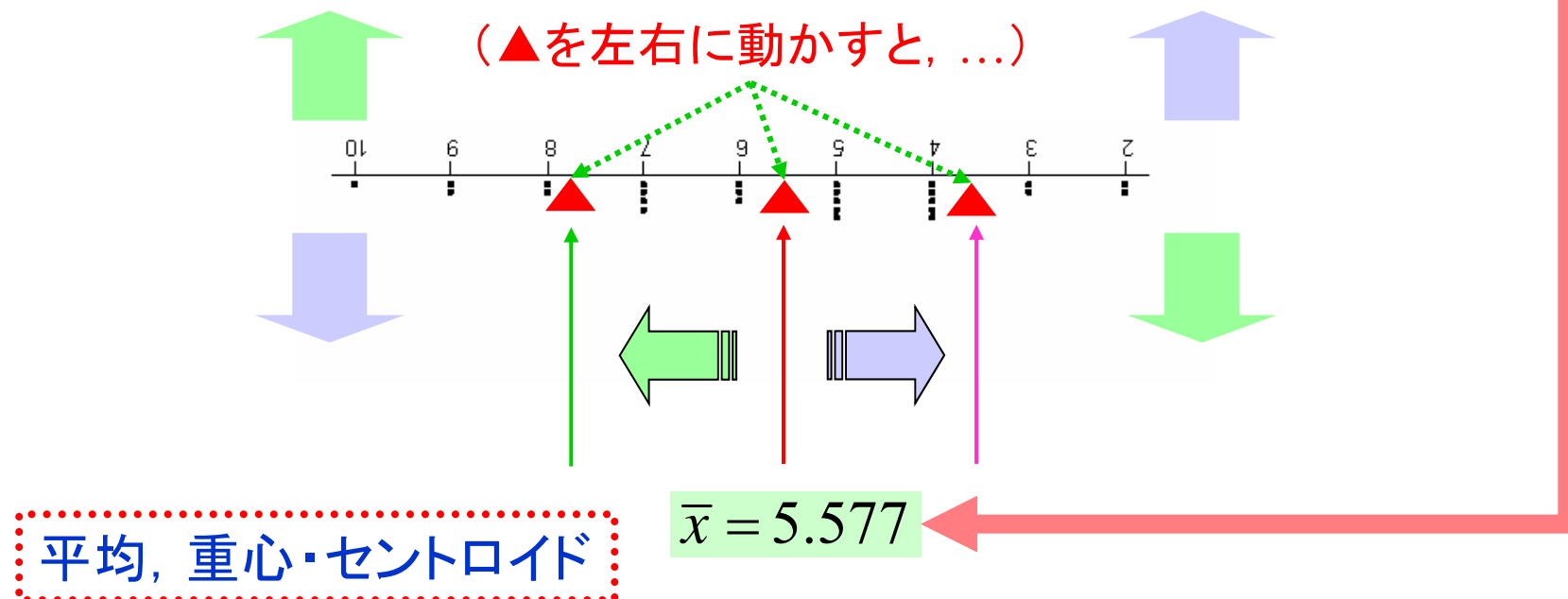
平均のまわりの偏差の和と平均(重要)

- 個々の偏差は個々の測定値の変動を示してはいる.
- しかし, 偏差の和や平均とするとゼロとなる.
- 平均のまわりの偏差の和や平均では, 測定値の示す分布の変動は測れない.
- どのようなデータでも, そのチラバリ(変動)の大小に関係なく常にゼロとなる.
- 個々のデータが, 平均からどのように離れているかに無関係である.
- 平均が重心となっているからなり立つ関係である.

平均の別の見方(重要!!)

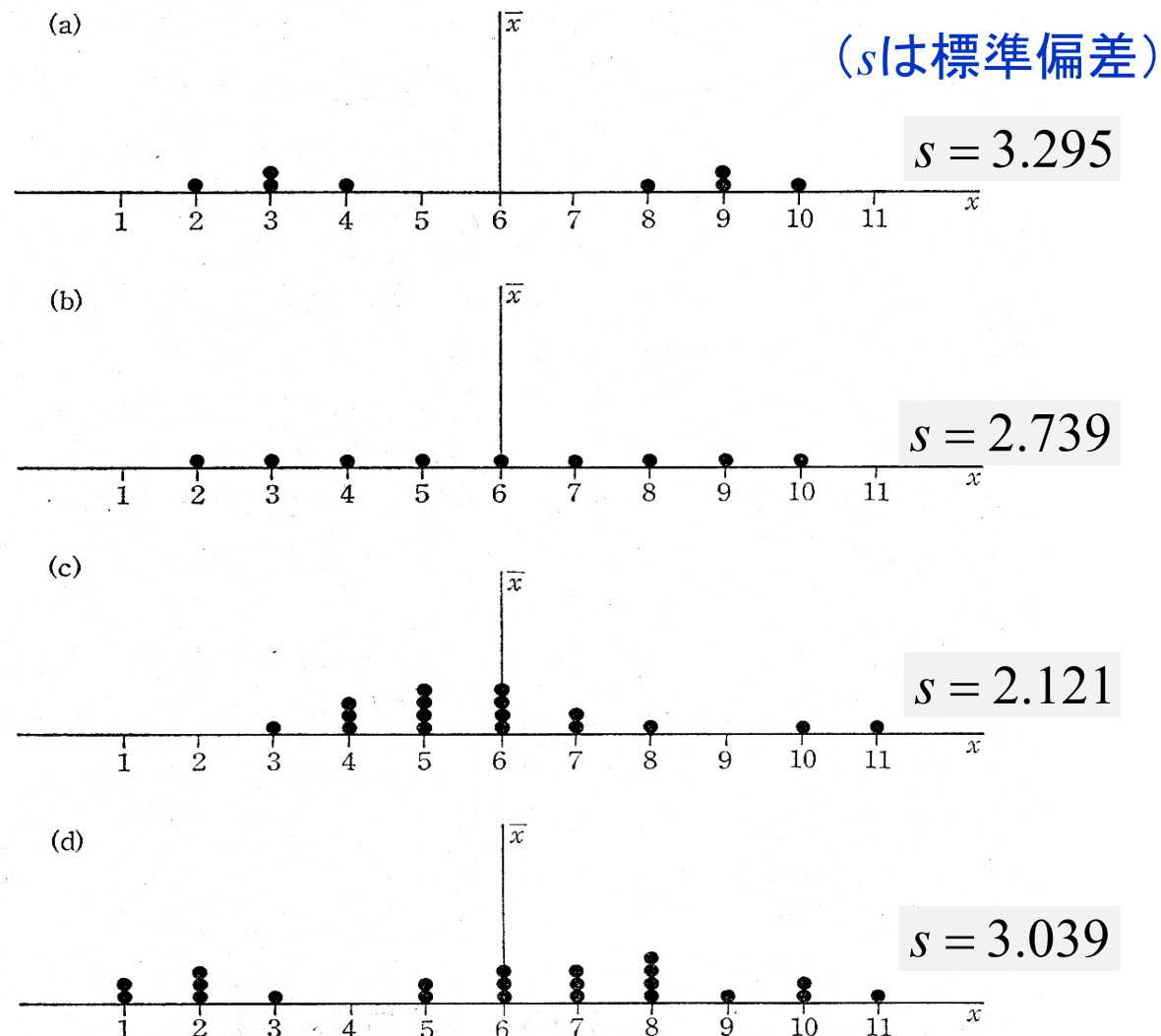
- 平均は**重心**(セントロイド; centroid)ともいう.
- 物理的な意味での“**重心**”のこと.
- 竿秤「やじろべえ」を考える⇒**支点**でバランスをとる.

各点(データ)を均質で等しい重さの「おもり」としよう
バランスがとれる位置はどこか? ⇒ 物理の**動的モーメント**を考える
バランスをとる位置 ⇒ 左右のモーメントが**ゼロ** ⇒ 動かない



平均が同じで変動(バラツキ)が異なる例

“位置(平均)は
同じだが形
(変動)は異なる”と解釈する。
点のチラバリ方
は違う。
どう区別する？



平均は同じ、標準偏差は異なる例
測定している「分布の特徴」が異なる

観察すると, ...

- 平均は単なる位置の目安, これだけでは測定値の変動(バラツキ)は分からない.
- この例は平均はすべて「6」である.
- 平均が分布を正確に代表しない例である.
- 平均という縮約値に変えたことで情報の損失がある.
(縮約による情報の損失)
- しかし分布の形(変動, バラツキ)は異なる.
- 図に標準偏差「 s 」を書き入れてある. これは異なる. これは何を測っているのか.

平均, 変動・バラツキ, 偏差, 情報の損失

③「偏差平方和」あるいは「平方和」

- 偏差平方和は“平方和”ともいう.
- まず“ある値 a のまわりの偏差平方和”を考える.
- 値 a と個々の測定値 (x_i) との 距離の二乗の和 (平方距離の和) に相当.
- つまり“変動”を測っている. 遠い・近いとその広がり具合 (バラツキ) を測る.

a のまわりの偏差平方和
あるいは a のまわりの平方和

$$S^* = \sum_{i=1}^n (x_i - a)^2 (\geq 0)$$

平方和 (s.s.: sum of squares)

とくに、平均のまわりの「平方和」

- “平均のまわりの平方和” (偏差平方和)を考える.
- 平均 (\bar{x}) と個々の測定値 (x_i) との 距離の二乗の和 (平方距離の和) に相当する.
- 平均のまわりの変動 を測っている. 測定値の遠い・近いとその 拡がり具合 (バラツキ) を測る.

平均のまわりの偏差平方和

あるいは平均のまわりの平方和

$$S = \sum_{i=1}^n (x_i - \bar{x})^2 (\geq 0)$$

- このままでは測定値の大きさ(n)が異なる集団は、比較できない.
- ここでまた偏差平方和の“平均”を作る.
- これを“分散”(variance)という(K. ピアソンが名付けたとされる).
- 標本から得た測定値から求めると“標本分散”という.
- 分散とは測定値と平均との“距離の二乗の平均”と言い換えられる.
- 個々の測定値の平均からの離れ具合を測る. 測定値間の“違いを識別”する情報量と考える.
- パターン認識でいう“特徴抽出”の一種とみてよい.

④「分散」(標本分散, 不偏分散)を作る

標本分散 あるいは “分散”

$$s_0^2 = \frac{1}{n} S = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

注: n が大きくなれば, $s_0^2 \doteq s_u^2$ となる.

$$\frac{1}{n} \sum_{i=1}^n (\bullet)^k$$

この形と括弧内に注意

“不偏分散”(不偏な標本分散のこと)

$$s_u^2 = \frac{1}{n-1} S = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

分散 (variance), 標本分散

不偏分散 (不偏な標本分散; unbiased variance)

⑤「標準偏差」を作る

- 分散は測定単位の二乗になる. 原単位と比較できない.
- 分散の正の平方根とし, 元の測定値の単位にそろえる.

標本標準偏差 あるいは “標準偏差”

$$s_0 = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

標準偏差

$$s_u = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

標準偏差 (S.D.: standard deviation)

平均の性質(重要)

- 平均は“はずれ値”の影響を受けやすい.
- 平均から遠い位置にある測定値に影響される.
- 典型的な例.
 - 例1: 世帯所得金額の分布
 - 例2: ウェブ調査の回答所要時間の分布
- 中央値(メジアン), トリム平均などを併用する.
- 対応分析で得る成分スコアの平均も同様に考える.

分散と標準偏差の性質(重要)

- 分散, 標準偏差のいずれも, 平均から遠い位置にある測定値の影響を受ける.
- 分散の式で, 偏差平方となっているから二乗の大きさで, 大きい値の影響を受ける.
- つまり, “はずれ値”の影響を受けやすいこと.
- 平均と分散あるいは標準偏差という統計量だけでは“分布の特徴”は測れない(分布の特徴の一部を観察しているだけ).
- 他の統計量の観察も必要ということ.

平均や平方和・分散を用いる理由は？

- なぜ、平方和や分散に“平均が入っている”のか。
- 他の値ではいけないのか。

平均

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

平方和

$$S = \sum_{i=1}^n (x_i - \bar{x})^2$$

分散
(標本分散)

$$s_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

★応用問題：解答は以下を解くことで分かる

Q1: つぎの値を最小にする a は何か

$$Q(a) = \sum_{i=1}^n (x_i - a)^2 \geq 0 \quad (a \neq 0) \Rightarrow (\text{最小化})$$

Q2: つぎの関係が常に成り立つことを示せ
(平均値のまわりの平方和がもっとも小さいこと)

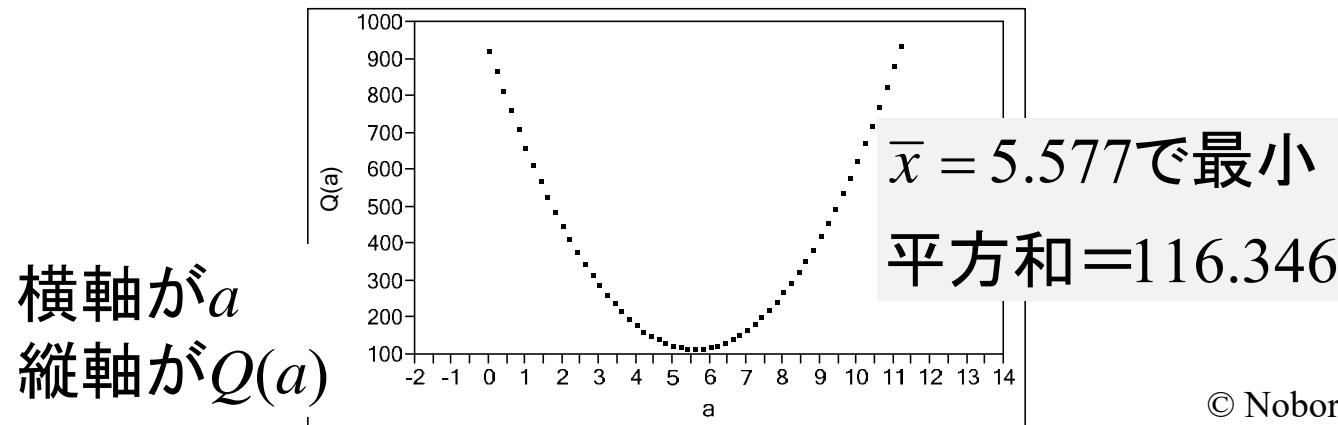
$$\sum_{i=1}^n (x_i - \bar{x})^2 \leq \sum_{i=1}^n (x_i - a)^2 \quad (a \neq 0)$$

以下となる

$$Q(a) = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - a)^2 \geq 0$$

ここで、 x_i, \bar{x}, n には、実際は測定値から得た数値
が入っている。よって、 $Q(a)$ は、 a を未知数とする関数。

そして、“ a =平均”のときにこの a の2次関数は“**最小**”となる、
つまり“**平方和**”となる。 $n=26$ のデータで確かめる(下図)。



要約: 平均, 平方和と分散の関係

- 平均は**重心**である, つまり測定値の分布の**均衡点**となっている.
- 動的モーメントがない, **エネルギー最小**の位置, つまりヤジロベエが均衡をとった状態(重心).
- 重心ではあるが, かならずしも, その平均のまわりにデータが集中・分布することでは**ない**.
- 単なる分布の**位置の目安, 代表値**であること.

(つづき)

- (平均のまわりの) 平方和は平均値と相対的に均衡をとった関係にある.
- 平方和あるいは分散も, 竿秤が均衡となったことを意味する (変動が最小, 動的モーメントがゼロ).
- 平均と平方和, 分散あるいは標準偏差は連動していると考えるのが自然である.
- 一般に, 平均値 (位置) が変われば平方和, 分散 (変動, 形) も変わる. 注: 例外は“正規分布”のみ.

(つづき)

- かりに平均値や分散が同じであっても“分布は同じ”とは限らない.
- この2つの統計量(平均, 分散または標準偏差)だけでは分布を完全に説明できない.
- さらに高次の積率の利用を考える(例: 歪度, 尖度).
- あるいは別の統計量も用いる(例: 順序統計量の中央値, 範囲, 四分位範囲など).
- つまり, 複数の統計量を用いて総合的に測定値の“分布”を評価すること. それが“探査”ということ.

「標準化」とは？

- 測定値(データ)に加減乗除を行う操作の1つ(一次変換).
- とくに, 平均と標準偏差を用いる変換を“標準化”という.
- まず, 以下を再確認する.

測定値 $x_i \ (i = 1, 2, \dots, n)$

平均 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

標本分散 $s_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

標準偏差 $s_0 = \sqrt{s_0^2}$

標準化による測定値の変換

$$z_i = \frac{x_i - \bar{x}}{s_0} \quad (i = 1, 2, \dots, n)$$

- 個々の測定値から平均を引いて標準偏差で割る.
- これを“標準化”という.
- 言い換えると, [平均値のまわりの偏差] ÷ [標準偏差]のこと.
- 標準偏差で割らなければ, 分布の形は変わらず位置が変わるだけ.

標準化 (standardization, normalization)

標準化変数の平均と分散・標準偏差

変数 z_i の平均 \bar{z} , 分散 s_z^2 , 標準偏差 s_z を求める.

$$\text{平均: } \bar{z} = \frac{1}{n} \sum_{i=1}^n z_i = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_0} \right) = \frac{1}{ns_0} \sum_{i=1}^n (x_i - \bar{x}) = 0$$

$$\text{分散: } s_z^2 = \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2 = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_0} \right)^2$$

$$= \frac{1}{s_0^2} \overbrace{\left\{ \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right\}}^{=s_0^2} = 1$$

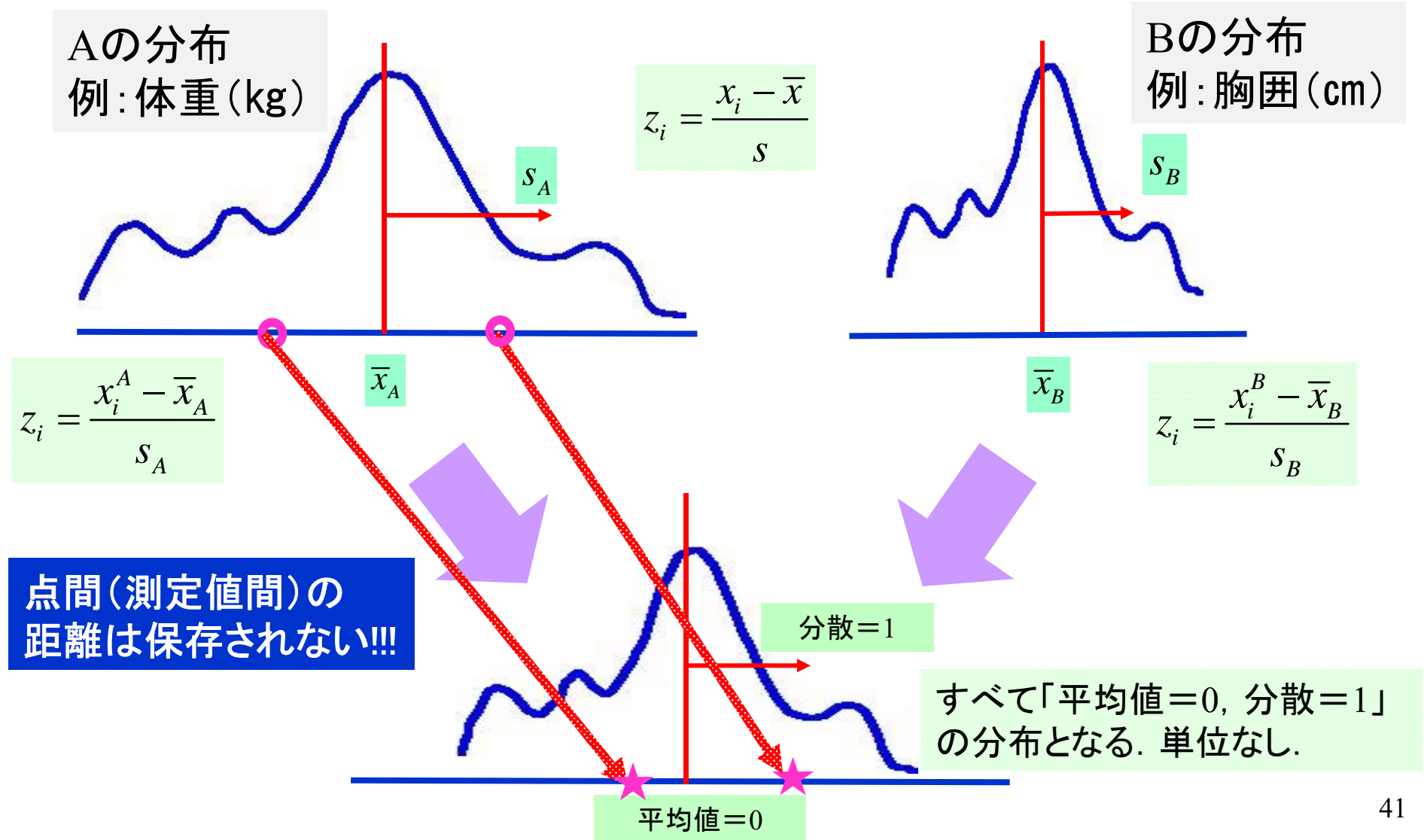
- 標準化変数の“平均は0”となる.
- 標準化変数の“分散は1”となる. 標準偏差も1となる.

(つづき)

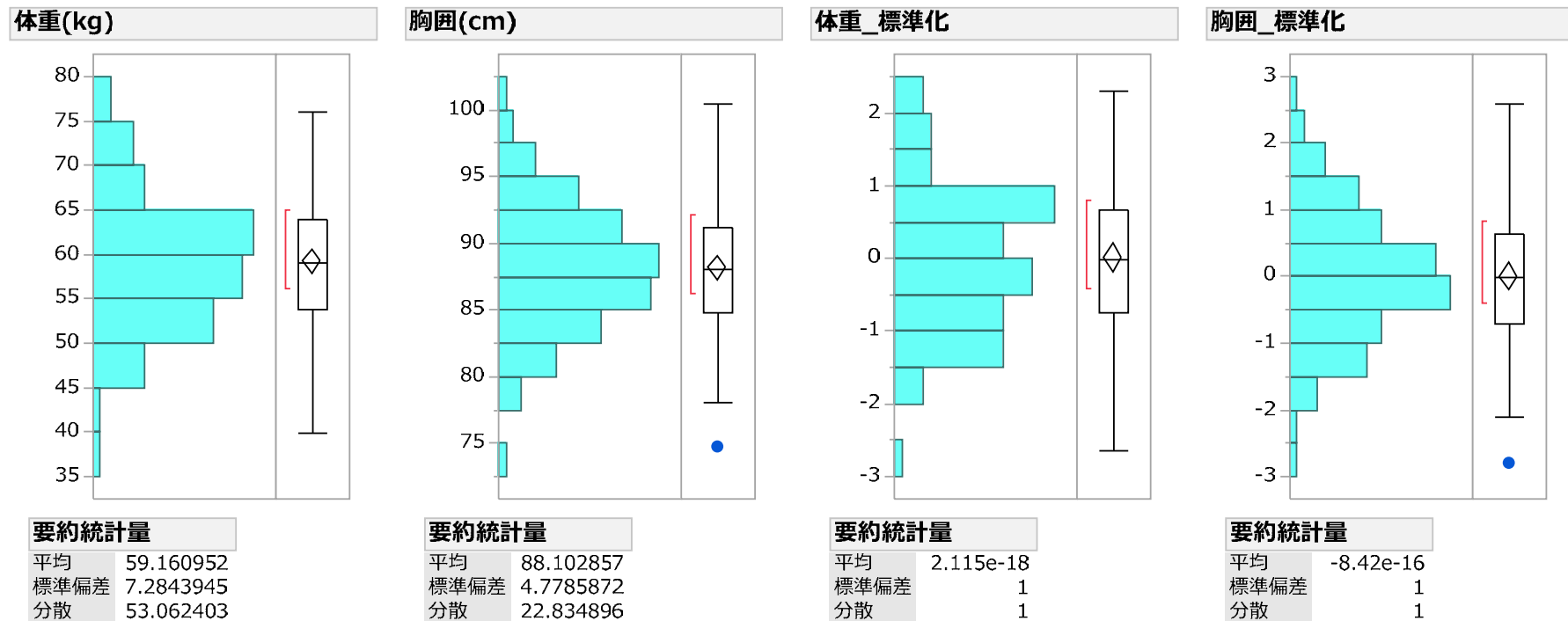
- 測定値の分布の平均値を0, 分散・標準偏差を1とする操作を“標準化”という.
- 標準化の操作はさまざまな場面で形を変えて登場.
- たとえば, 相関係数の誘導, 対応分析, 主成分・因子分析の“スコア”の標準化などなど.
- 式の形から, 単位がない(無名数である).
- z スコアあるいは標準化スコアということがある.
- 「正規分布」には直接関係しない単なる変数変換(一次変換).

注: 元の変数が正規分布なら, 標準化変数も正規分布.

標準化を図で考えると, ...



「標準化」を数値例で確認



体重(kg)

胸囲(cm)

無名数(単位なし)

- 平均は、わずかの計算誤差が生じ、指数ベキ表示になっているが「ゼロ」。分散と標準偏差は「1」になる。

標準化の意味(重要 !!!)

- ①位置と尺度の変換:どんな測定値の元の分布の特徴(位置と尺度の情報)も「**平均値=0, 分散=1**」の分布とする.
 - ②測定単位の消去:平均値と標準偏差による一次変換で単位はない(**無名数**).たとえば特性の元の測定単位が「cm」であると,平均値と標準偏差の単位は同じcmだが,標準化後の単位はない.
 - ③異なる特性間の比較:単位がないので,異なる測定単位の特性間の比較が可能.たとえば,単位が「cm」と「kg」との測定データ間の比較が可能(ただし,標準化変数の意味で).相関係数はこの性質を利用.
 - ④標準化の影響:標準偏差で除し尺度を変えたので,個々の測定値間の距離は保存されない.元の測定値が示す分布とは形が異なる分布を観察していることに注意.
- ◎とくに④には注意. 成分スコアの解釈時に注意.

付録1：主な統計量とグラフィカル表現法

目 的	積率型統計量 から得る統計量	順序統計量から得る統計量 (積率で表記されない統計量)
位置の指標	平均, 割合	中央値(メジアン), モード(最頻値) 中点値(ミッドレンジ)
変動・バラツキの指標	平方和と分散 標準偏差 絶対偏差と平均偏差	範囲, 四分位範囲(ヒンジ散布度) 中央絶対偏差(MAD) 最短の半分(SHORTH)
歪み・偏りの指標	歪度(歪み)	上の一部が利用可能
尖り・集中度の指標	尖度(尖り)	
用いるグラフィカル表現法	ヒストグラム ドットプロット 箱ひげ図 幹葉図	箱ひげ図 パーセンタイル・グラフ ビジネス・グラフ (棒グラフ, 折れ線, 円グラフなど)

主な記述的統計量(1)

目 的	統計量	用途, 留意点
位置の指標 (代表値)	平均 割合・比率	分布の平均的位置, 代表値 割合(比率)(ベルヌーイ試行型, 二項選択的)
	中央値 (メジアン)	<u>歪んだ分布, はずれ値のある分布に対して頑健</u> (これをロバストという) 例: 所得分布, Web調査回答所要時間分布, ログ解析の評価など
	中点値 (ミッドレンジ)	分布のおよその位置の見当をつける 他の統計量と比較するのがよい <u>測定値の数が少ないときに有効, 便利</u>
	モード (並数, 最頻値)	度数分布のとき, 最大の度数(頻度)をもつ階級の値 度数分布の階級幅が一定で“単峰型”の分布に用いるのがよい

主な記述的統計量(2)

目 的	統計量	用途, 留意点
変動・バラツキ の指標 (形の指標)	平方和と分散	変動を測る統計量, いろいろな意味で <u>非常に重要</u> <u>はずれ値の影響を受けやすい</u> 測定単位の2乗であること／平均のまわりの <u>2次積率</u>
	標準偏差	変動を測る基本的な統計量 はずれ値の影響を受けやすい 測定単位に揃えてある
	平均偏差	変動を測る指標 偏差の絶対値を使う. はずれ値の影響が <u>緩和される</u> 積率として一般化できない
	範囲(レンジ)	簡単に測れる変動の指標 測定値の数が少ないとき, 標準偏差の代用となりうる (例: 管理図での応用)
	四分位範囲 (ヒンジ散布度) 最短の半分(SHORTH) 中央絶対偏差(MAD)	はずれ値の検出の目安 他の分位数との併用 分布の変動の簡易指標として便利

主な記述的統計量(2)

目 的	統計量	用途, 留意点
歪み, 偏り	歪度(歪み)	<p>測定値の分布の<u>歪みや偏りの探査</u></p> <p>平均のまわりの<u>3次の積率</u> (単位が, 元の3乗となる)</p> <p>その変形, とくに標準化し無名数化した歪度 (正規分布のときに0と調整)</p> <p>はずれ値の影響を受けやすい</p>
尖り, 集中度	尖度(尖り)	<p>測定値の<u>尖りや集中度の探査</u></p> <p>平均のまわりの<u>4次の積率</u> (単位が, 元の4乗となる)</p> <p>その変形, とくに標準化し無名数化した尖度 (正規分布のときに0と調整)</p> <p>はずれ値の影響を受けやすい</p>

たとえば, JMPの要約統計量オプションは,

<input checked="" type="checkbox"/> 平均[デフォルトはオン]	<input checked="" type="checkbox"/> 最大値
<input checked="" type="checkbox"/> 標準偏差[デフォルトはオン]	<input checked="" type="checkbox"/> 中央値
<input checked="" type="checkbox"/> 平均の標準誤差[デフォルトはオン]	<input type="checkbox"/> 最頻値
<input checked="" type="checkbox"/> 平均の上側信頼区間[デフォルトはオン]	<input type="checkbox"/> トリム平均
<input checked="" type="checkbox"/> 平均の下側信頼区間[デフォルトはオン]	<input type="checkbox"/> 幾何平均
<input checked="" type="checkbox"/> N[デフォルトはオン]	<input checked="" type="checkbox"/> 範囲
<input type="checkbox"/> 重みの合計	<input checked="" type="checkbox"/> 四分位範囲
<input checked="" type="checkbox"/> 合計	<input type="checkbox"/> 中央絶対偏差
<input checked="" type="checkbox"/> 分散	<input type="checkbox"/> ロバスト 平均
<input checked="" type="checkbox"/> 歪度	<input type="checkbox"/> ロバスト 標準偏差
<input checked="" type="checkbox"/> 尖度	<input type="checkbox"/> α 水準の設定 <input type="text" value="0"/>
<input checked="" type="checkbox"/> 変動係数	<input type="checkbox"/> トリム平均のパーセントを設定 <input type="text" value="0"/>
<input checked="" type="checkbox"/> 欠測値 N	<input type="checkbox"/> すべての最頻値を表示
<input type="checkbox"/> ゼロの個数	
<input type="checkbox"/> 一意な値の個数	
<input checked="" type="checkbox"/> 無修正平方和	
<input checked="" type="checkbox"/> 修正平方和	
<input type="checkbox"/> 自己相関	
<input checked="" type="checkbox"/> 最小値	

- このリストにある項目はほとんどが統計量と考える.
- デフォルト以外に設定条件を自分で指定できる.
- それぞれが何を意味し, 何のために用いるかを知る.

付録2: 標本分散と不偏分散の関係

- 標本分散 s_0^2 と不偏分散 s_u^2 の関係を知っておくことは重要.
- つねに標本分散のほうが不偏分散よりも小さい.
- 測定値の大きさ(n)が大きくなればほぼ同じ($n \geq 50 \sim 100$).

$$s_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \Rightarrow \sum_{i=1}^n (x_i - \bar{x})^2 = n s_0^2$$

\Downarrow

$$s_u^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n}{n-1} s_0^2$$

$$s_0^2 = \frac{n-1}{n} s_u^2 = \left(1 - \frac{1}{n}\right) s_u^2 < s_u^2$$

$\therefore 1 - \frac{1}{n} \rightarrow 1 \text{ (} n \rightarrow \infty \text{ のとき)}$

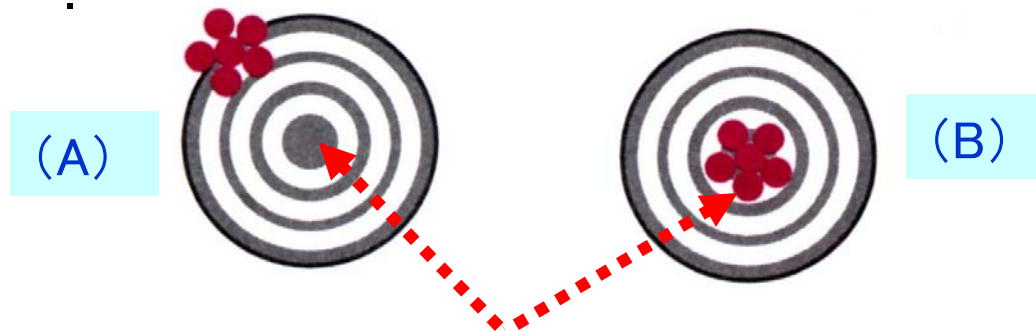
① s_u^2 のほうが統計量として s_0^2 よりも母集団分散(母分散) σ^2 をあてやすい“不偏な(unbiased)”推定量となっている.

② つまり s_0^2 には“偏り(bias)がある.

- これらの「正確さ」と「精度」は, 射撃と的の関係で考えると 分かりやすい.
- とくに, 偏りは「正確さ」に関わること.

母分散(σ^2)を正確に当てたいとする

- 的の真ん中が当てたい母分散だとする.
- 標本分散では, (A)のようになり, “偏り”がある.
- 一方, 不偏分散を用いると, (B)のように“偏りなく”母分散を当てられる. つまり, “不偏”となる.
- 標本分散と不偏分散の関係式から, 不偏分散は $(n-1)$ で割る必要がある(そうしないと, 不偏にならない).
- ここは偏りなく当たるかどうかを考え, その精度までは考えていない(精度: 下の図の赤い点のチラバリの大きさ)



(母分散) σ^2 は的の中央とする.

JMRAマーケティング・リサーチ講座

質的データのマイニング のための対応分析法

スライド資料[その2]

大隅 昇

ohsumi@ss.ij4u.or.jp

<http://wordminer.org/>

Copyright by Noboru Ohsumi

この資料[その2]の内容

- ここで、対応分析法で重要な役割を果たす“**ピアソンのカイ二乗統計量**”(χ^2 統計量)について述べる.
- K. Pearson(父ピアソン)が提案した.
- 非常にすぐれた発想で、実用的にも有用である.
- 近代統計学の中でも顕著な成果の1つ.
- 以下、断りなければ“カイ二乗統計量”はこの統計量を指すとする. [いろいろなカイ二乗統計量がある]
- ベンゼクリもこの統計量がなければ、(彼の言う)データ解析はなりたたないとしている.

χ^2 統計量(chi-squared statistic)

K. Pearson(1857-1936)の息子も著名な統計家; E.S. Pearson

(つづき)

- 分布あるいはモデル間のある距離(乖離度)を測る統計量の1つ.
- 関連研究が無数にある. とくに, 質的データの分析における多数の成果がある.
- あのR.A. Fisher(フィッシャー)も, 分割表の研究で, さまざまな成果を挙げている. 例: 再生公式(推移公式)
- また, カイ二乗統計量の自由度を巡るピアソンとの有名な論争もある. [あとですこし述べる]

R.A. Fisher(フィッシャー): ピアソンと同時代に多大の業績をあげた統計学界の偉人・巨人(最尤法, 実験計画法など), あとで登場.

カテゴリカル・データ分析は, こうした本に詳しい:

A. Agresti (1990, 2010): *Categorical Data Analysis*, John Wiley & Sons.

(つづき)

- 説明の導入部として、ある調査データから得た“クロス表”に対応分析法を適用，簡単な「数値例」を調べる.
- とくに，何が質的データで，数量化（数量化得点，成分スコア）とは何を行うのか，の基本要素を確認する.
- これから登場する基本的な用語句（記号，記法）の解釈や意味を知る.
- 伝統的な統計科学の慣用語句のほかに，対応分析法に固有の（ベンゼクリ流の）“方言”が混ざるので注意.

カイ二乗統計量による検定

- ほとんどの統計関連書に登場する統計的検定の基本的な方法の1つ.
- 分布間の距離(乖離度)を測る指標の一つと考えられる.
- (分布の)”適合度検定“や分割表・クロス表の”独立性の検定“に用いる.
- 注意として“離散型確率分布”であるカイ二乗統計量が, “連続型確率分布”の χ^2 分布に“(よく)近似”することを用いて検定を行うこと.
- よって, χ^2 分布の知識も必要となる(数表のひき方など).

(つづき)

- カイ二乗統計量の特徴、使い方の要点を知るため、対応分析法の導入部として必要なことのみ触れる(細かい数理的なことは略す)。
- 基礎情報として、2つの使い方について述べる。
 - ①「適合度の検定」、分布の同等性などを調べるとき。
 - ②「独立性の検定」、分割表・クロス表の2項目間の関連性を調べるとき。
- カイ二乗統計量は応用範囲が広く、離散型確率分布のほとんどの分野で登場する。
- 多元クロス表分析の“対数線型モデル”などでも多用する。

ピアソンのカイ二乗統計量

- 大まかに(一般的に)記すと, カイ二乗統計量とは以下のような統計量のこと.
- 和の記号(Σ)は, 考えられる事象すべての組合せ(通り数)の意味. うしろで例でみる.

$$\chi_p^2 = \sum_{\left(\begin{smallmatrix} \text{すべての} \\ \text{組合せ} \end{smallmatrix}\right)} \frac{(\text{実現度数} - \text{期待度数})^2}{\text{期待度数}}$$

$$\chi_p^2 = \sum_{\left(\begin{smallmatrix} \text{すべての} \\ \text{組合せ} \end{smallmatrix}\right)} \frac{(\text{観測データの情報} - \text{モデルの情報})^2}{\text{モデルの情報}}$$

参考: 尤度比カイ二乗統計量とは $\chi_L^2 = 2 \sum_{\left(\begin{smallmatrix} \text{すべての} \\ \text{組合せ} \end{smallmatrix}\right)} (\text{実現度数}) \times \ln \left(\frac{\text{実現度数}}{\text{期待度数}} \right)$

期待度数と実現度数

- 基本は、分布間のある種の距離を考えていること.
- 式の分子は平方距離になっている. それを分母(期待度数, モデル)で割っている点に特徴がある.
- 測定で得た“**実現度数**”(観測度数)と, 理論上の度数つまり想定したモデルから得られる(推定した)“**理論度数**”との間の乖離を測る.
- (帰無)仮説を「モデル」(理論度数)とすると, これに不適合であるほど統計量の実現値が大きくなる, つまり距離が大きい, となる(はず).
- 検定操作を通じて, 両者(実現度数と理論度数)が近い
か遠いかを測る.
- まず, もっとも簡単な“**適合度の検定**”を考える.

「適合度の検定」の簡単な例

- 簡単な例として「サイコロ投げ」実験で、サイコロの目の出方が均等であるかどうかを調べる。[サイコロは正6面体とする]。
- 目の出方が一様であるかを“実現度数”（観測度数）を用いて検定すること。
- “帰無仮説”とする均等モデル（理論分布）が与える（モデルが予想する）“期待度数”とのある種の距離を測る問題。
- 重要なキーワードは“実現度数”，“期待度数”である。
 - 実現度数＝実験で得られる実際の観測度数，経験分布ということ。
 - 期待度数＝モデルとする理論分布から得た（推定した）度数。

帰無仮説

実現度数（経験分布）と期待度数（理論分布）

モデルの定式化

- 一般に, 2つの分布, つまり実現度数の経験分布と, 理論度数を与える理論分布の関係を図のように考える.
- k 個の層(セル)について, いくつかの観測値が実現値として得られたか, サイコロ投げならば「 $k=6$ 」で各目の出方に相当.

	A_1	A_2	\cdots	A_i	\cdots	A_k
実現度数	f_1	f_2	\cdots	f_i	\cdots	f_k
理論度数	e_1	e_2	\cdots	e_i	\cdots	e_k

←実際に手にした実現度数

←モデル・仮説から推定した期待度数

$$\sum_{i=1}^k f_i = \sum_{i=1}^k e_i = N \quad (\text{総度数})$$

ここで A_i ($i = 1, 2, \dots, k$) は互いに排反的

記号の用意と考え方(抜粋)

- ここでは, こんなふうに考えるのだ, としておく.
- “何を行っているのか”, “解釈”とその限界を理解することが重要.

帰無仮説 $H_0 : p_1 = p_1^*, p_2 = p_2^*, \dots, p_i = p_i^*, \dots, p_k = p_k^*$

対立仮説 $H_1 : H_0$ がなり立たない(上記の等号のいくつかが不成立)

$$\sum_{i=1}^k f_i = \sum_{i=1}^k e_i = N \quad (\text{総度数})$$

実現度数 f_i : 実際に観測した第*i*セル ($i = 1, 2, \dots, k$) の実現度数

期待度数 $e_i : e_i = Np_i^* \quad (i = 1, 2, \dots, k)$

注: 各セルが, 上のような割合でおこるとしたら, というモデル.

記号, 記法の用意と手順

$$\chi_p^2 = \sum_{i=1}^k \frac{(f_i - e_i)^2}{e_i} = \sum_{i=1}^k \frac{(\text{実現度数} - \text{期待度数})^2}{\text{期待度数}} \approx \chi_{k-1}^2$$

(χ_p^2 を自由度 $d.f. = k - 1$ の χ_{k-1}^2 分布で近似し検定する)

- ①片側検定となる.
- ②測定値から得た検定統計量の実現値 χ_{k-1}^2 を有意水準 α の χ^2 分布の限界値 $\chi_{k-1}^2(\alpha)$ と比べ, $\chi_{k-1}^2 \geq \chi_{k-1}^2(\alpha)$ ならば帰無仮説 H_0 を棄却し有意とする.
- ③そうでなければ, 帰無仮説は棄却せず, 有意でない, とする.

離散分布のカイ二乗統計量(Σ で書いている)を連続分布の χ^2 分布で“近似”することに注意する.
「自由度」とはなにか(後述).

例：簡単な数値例

【問題】

あるサイコロを240回 ($N=240$) 投げたところ表のようになったという(実現度数 f_i の欄). このサイコロの目の出方は均等といってよいかを検定で確かめよ.

- 等比率・等頻度分布の検定となる. [適合度の検定]
- 帰無仮説は「各目の出方は均等, $1/6$ である」とする.
- ここではこうすること. モデルはいろいろある(前述).
- 2つの分布が近いか遠いかを測るので片側検定.

(つづき)

	A_1	A_2	A_3	A_4	A_5	A_6	
サイコロの目	1	2	3	4	5	6	(総和)
実現度数(f_i)	41	32	50	42	43	32	240
期待度数(e_i)	40	40	40	40	40	40	240

- 均等である必要はない, 各目の出方がある割合になるかをみる.
- 「適合度の検定」という. ここでは均等とするモデル, という事.

帰無仮説, 期待度数, 統計量の算出

帰無仮説 $H_0 : p_1 = p_2 = \cdots = p_i \cdots = p_6 = \frac{1}{6}$ (目の出方が均等)

期待度数 $e_i : e_i = Np_i = 240 \times \frac{1}{6} = 40 (i = 1, 2, \cdots, 6)$
(各セルの期待度数)

実際にカイ二乗統計量の実現値を“この1組”の観測値から求める. これは離散的情報の和. ある種の距離.

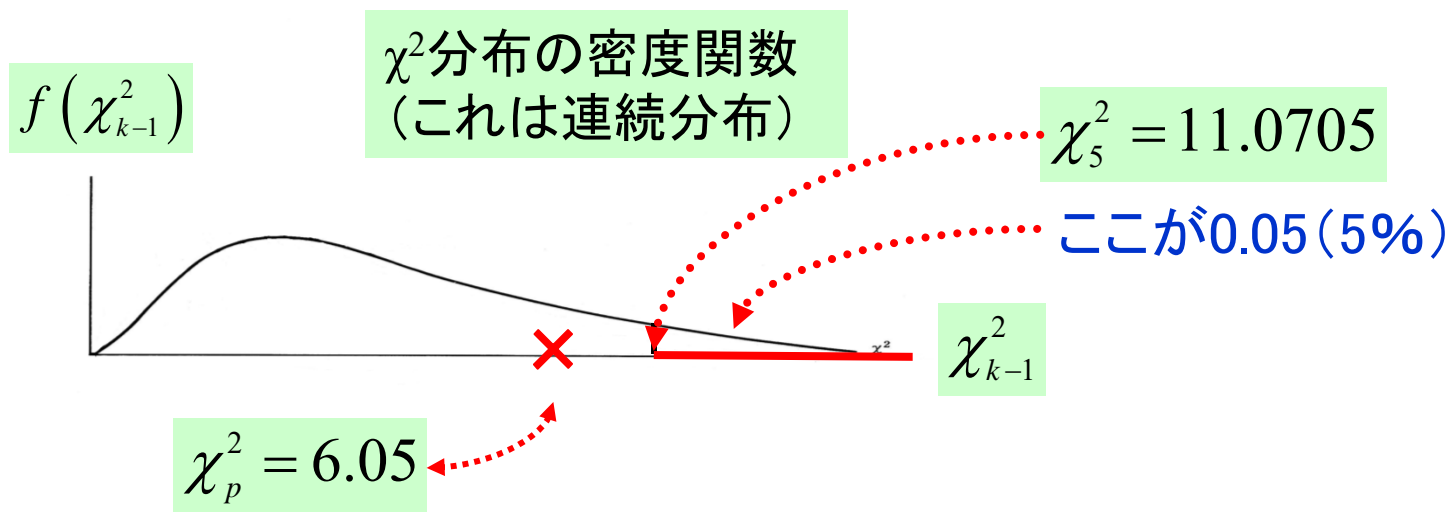
$$\begin{aligned}\chi_p^2 &= \sum_{i=1}^k \frac{(f_i - e_i)^2}{e_i} = \frac{(41 - 40)^2}{40} + \frac{(32 - 40)^2}{40} + \frac{(50 - 40)^2}{40} \\ &\quad + \frac{(42 - 40)^2}{40} + \frac{(43 - 40)^2}{40} + \frac{(32 - 40)^2}{40} = 6.05\end{aligned}$$

検定の手順と結果解釈

$$\chi_p^2 = \sum_{i=1}^k \frac{(f_i - e_i)^2}{e_i} \approx \chi_{k-1}^2$$

$$\chi_{k-1}^2(\alpha) = \chi_5^2(0.05) = 11.0705 \text{ (有意水準5\%に対する棄却点)}$$

$$\underbrace{\chi_p^2 = 6.05}_{\text{実現した値}} < \underbrace{\chi_5^2(0.05) = 11.0705}_{\text{有意水準5\%に対する棄却点}}$$

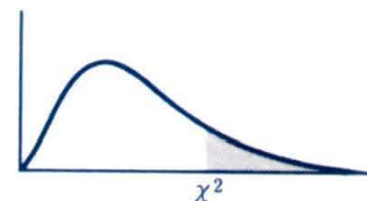


(つづき)

- 得られたカイ二乗統計量(離散値)の実現値を(連続型確率分布である) χ^2 分布の上で比べる(近似).
- ここで, 有意水準は0.05(5%)とする(指定).
- χ^2 分布の確率表から, 自由度 $d.f.=k-1=5$, 有意水準 $\alpha=0.05(5\%)$ に対する棄却点を求める.
- ここでは「**有意でない**」となり, 帰無仮説を棄却しない, となる(消極的に採択する).
- 「均等でないとはいえない」(二重否定)と読む(背理法的な解釈).
- **解釈**: サイコロの目の出方は均等であるらしい(「ある」と断定はできない). 5%のリスクがある(有意水準).

カイ二乗分布の確率表の例

表側の数字は自由度 (ν) を表わす。表頭の数字は χ^2 の値が表の中の数値を超える確率 (P) を表わす。 $\nu > 100$ の場合には、 $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ が近似的に標準正規分布に従うことを利用して正規分布表を用いればよい。



$P \backslash \nu$	0.995	0.975	0.050	0.025	0.010	0.005
1	0.0 ⁴ 3927	0.0 ³ 9821	3.84146	5.02389	6.63490	7.87944
2	0.010025	0.050636	5.99147	7.37776	9.21034	10.5966
3	0.071721	0.215795	7.81473	9.34840	11.3449	12.8381
4	0.206990	0.484419	9.48773	11.1433	13.2767	14.8602
5	0.411740	0.831211	11.0705	12.8325	15.0863	16.7496
6	0.675727	1.237347	12.5916	14.4494	16.8119	18.5476
7	0.989265	1.68987	14.0671	16.0128	18.4753	20.2777
8	1.344419	2.17973	15.5073	17.5346	20.0902	21.9550
9	1.734926	2.70039	16.9190	19.0228	21.6660	23.5893
10	2.15585	3.24697	18.3070	20.4831	23.2093	25.1882
11	2.60321	3.81575	19.6751	21.9200	24.7250	26.7569
12	3.07382	4.40379	21.0261	23.3367	26.2170	28.2995
13	3.56503	5.00874	22.3621	24.7356	27.6883	29.8194
14	4.07468	5.62872	23.6848	26.1190	29.1413	31.3193



★参考: 比率を用いるとき

- 調査データのように回答比率で与えられる場合もある. このときは, 統計量を下のように書き替えればよい.

$$\sum_{i=1}^k f_i = \sum_{i=1}^k e_i = N \quad (\text{総度数})$$

実現度数 f_i : 実際に観測した第*i*セル ($i=1, 2, \dots, k$) の実現度数

期待度数 e_i : $e_i = Np_i^*$ ($i=1, 2, \dots, k$)

ここで実現度数, 期待度数をそれぞれ比率に変換する.

$$\text{実現比率: } \hat{p}_i = \frac{f_i}{N} \quad (i=1, 2, \dots, k)$$

$$\text{理論比率: } p_i^* \quad (i=1, 2, \dots, k)$$

$$\begin{aligned} \chi_p^2 &= \sum_{i=1}^k \frac{(f_i - e_i)^2}{e_i} = \sum_{i=1}^k \frac{(N\hat{p}_i - Np_i^*)^2}{Np_i^*} \\ &= \sum_{i=1}^k \frac{N(\hat{p}_i - p_i^*)^2}{p_i^*} \approx \chi_{k-1}^2 \quad \left(\begin{array}{l} \chi_p^2 \text{を自由度 } d.f. = k-1 \text{ の} \\ \chi_{k-1}^2 \text{ 分布で近似し検定} \end{array} \right) \end{aligned}$$

ここで p_i^* は仮定する母集団の各セルの理論比率とする.

「独立性の検定」について

- 2元クロス表の行 I と列 J の2つの項目の“関連性の有無” (“関連のない”こと, 独立なこと)をピアソンのカイ二乗統計量で調べる.
- 2つの項目間に“あるモデル”を仮定して, それを検定することで, 評価を行う方法をピアソンが提案した.
- 2項目 I, J の観測で得た実現度数(f_{ij})と理論分布(独立モデル)から得た期待度数(e_{ij})間の距離を測る.
- クロス表を一般的に記述するための記法を用意する. これは対応分析法の説明でも必要となる.

ここで, “2元クロス表”と“分割表”とは, 同じ意味に用いる.
クロス表の“モデル”は独立モデルに限らず多数工夫されている.

寸法 $(m \times n)$ の(2元)クロス表: $F = (f_{ij})$ の構成

$I \backslash J$	1	2	...	j	...	n	行和
1	f_{11}	f_{12}	...	f_{1j}	...	f_{1n}	f_{1+}
2	f_{21}	f_{22}	...	f_{2j}	...	f_{2n}	f_{2+}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
i	f_{i1}	f_{i2}	...	f_{ij}	...	f_{in}	f_{i+}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
m	f_{m1}	f_{m2}	...	f_{mj}	...	f_{mn}	f_{m+}
列和	f_{+1}	f_{+2}	...	f_{+j}	...	f_{+n}	f_{++}

項目 I の周辺分布

$$I = \{1, 2, \dots, m\}$$

$$J = \{1, 2, \dots, n\}$$

[項目 I と項目 J の選択肢]

項目 I, J の同時分布

$$f_{++} = \sum_{i=1}^m \sum_{j=1}^n f_{ij} = N \text{ (総度数)}$$

記法, 見方に慣れること
テキスト, 8ページあたりから

項目 J の周辺分布

クロス表ほか，記号・記法の用意

$$I = \{1, 2, \dots, m\}, \quad J = \{1, 2, \dots, n\} \quad [\text{項目}I\text{と項目}J\text{の選択肢}]$$

「選択肢」としておく。カテゴリー，モダリティともいう。この“標識”を質的データと考えている。

$$\mathbf{F}_{m \times n} = (f_{ij}) \quad [\text{寸法が}(m \times n)\text{のクロス表}]$$

$$f_{ij} : (i, j)\text{セルの度数} \quad [\text{同時分布}]$$

$$f_{i+} = \sum_{j=1}^n f_{ij} \quad [\text{行和: 項目}I\text{の周辺分布}]$$

$$f_{+j} = \sum_{i=1}^m f_{ij} \quad [\text{列和: 項目}J\text{の周辺分布}]$$

(実現度数)

$$\sum_{i=1}^m f_{i+} = \sum_{j=1}^n f_{+j} = N \quad f_{++} = \sum_{i=1}^m \sum_{j=1}^n f_{ij} = N \quad \left[\text{総度数 (= 全対象者数)} \right] \quad 21$$

(つづき)

- 対応分析を考える基本情報となるので、各記号が図のどこに位置し、何を示すかを確認する.
- 前にみたクロス表, 同時分布, 行和, 列和と周辺分布の関係を想起する.
- これを元に次の確率行列 $\mathbf{P} = (p_{ij})$ を作る.
- ここで式が“何を意味するか”をイメージとして捉えておこう.
- 対応分析法では”別の言い方“をするので注意(後述).
例: 質量, プロファイルとその重心(セントロイド), など.

確率行列: $\mathbf{P} = (p_{ij})$ を作る

$$\mathbf{P}_{m \times n} = (p_{ij}) \quad [\text{寸法が}(m \times n)\text{の確率行列}]$$

$$p_{ij} = \frac{f_{ij}}{N} \quad (\text{セル}(i, j)\text{の確率})$$

$$p_{i+} = \frac{f_{i+}}{N} = \frac{\sum_{j=1}^n f_{ij}}{N} \quad (\text{列のセル}(i, +)\text{の確率})$$

$$p_{+j} = \frac{f_{+j}}{N} = \frac{\sum_{i=1}^m f_{ij}}{N} \quad (\text{列のセル}(+, j)\text{の確率})$$

(実現確率)

$$p_{i+} = \sum_{j=1}^n p_{ij}, \quad p_{+j} = \sum_{i=1}^m p_{ij}; \quad \sum_{i=1}^m p_{i+} = \sum_{j=1}^n p_{+j} = 1, \quad \sum_{i=1}^m \sum_{j=1}^n p_{ij} = 1$$

[行和=列和=1] (全確率=1)

★参考: さらに記号の用意(あとで必要)

$$\mathbf{P}_{m \times n} = (p_{ij}) \quad [\text{寸法が}(m \times n)\text{の確率行列}]$$

$$\mathbf{P}_I = \text{diag}(p_{i+})_{m \times m} = \begin{pmatrix} p_{1+} & & & \mathbf{O} \\ & p_{2+} & & \\ & & \ddots & \\ \mathbf{O} & & & p_{i+} & \\ & & & & \ddots \\ & & & & & p_{m+} \end{pmatrix} \quad [\text{寸法が}(m \times m)\text{の対角行列}]$$

$$\mathbf{P}_J = \text{diag}(p_{+j})_{n \times n} = \begin{pmatrix} p_{+1} & & & \mathbf{O} \\ & p_{+2} & & \\ & & \ddots & \\ \mathbf{O} & & & p_{+j} & \\ & & & & \ddots \\ & & & & & p_{+n} \end{pmatrix} \quad [\text{寸法が}(n \times n)\text{の対角行列}]$$

$$\mathbf{r}_{m \times 1} = (p_{1+}, p_{2+}, \dots, p_{i+}, \dots, p_{m+})^t$$

[行の平均ベクトル]

$$\mathbf{c}_{1 \times n} = (p_{+1}, p_{+2}, \dots, p_{+j}, \dots, p_{+n})^t$$

[列の平均ベクトル]

確率行列: $\mathbf{P} = (p_{ij})$ (大きさ $m \times n$) の構成

$I \backslash J$	1	2	...	j	...	n	行和
1	p_{11}	p_{12}	...	p_{1j}	...	p_{1n}	p_{1+}
2	p_{21}	p_{22}	...	p_{2j}	...	p_{2n}	p_{2+}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
i	p_{i1}	p_{i2}	...	p_{ij}	...	p_{in}	p_{i+}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
m	p_{m1}	p_{m2}	...	p_{mj}	...	p_{mn}	p_{m+}
列和	p_{+1}	p_{+2}	...	p_{+j}	...	p_{+n}	1

項目 I の
周辺確率分布

項目 I, J の
同時確率分布

項目 J の
周辺確率分布

確認: テキストから引用

表 5 □ 確率行列 \mathbf{P}_{IJ} と行および列の相対確率

		項 □ 目 □ J						
		1	2	...	j	...	n	
項 □ 目 □ I	1	p_{11}	p_{12}	...	p_{1j}	...	p_{1n}	$\mathbf{r}_{m \times 1} = \begin{pmatrix} p_{1+} \\ p_{2+} \\ \vdots \\ p_{i+} \\ \vdots \\ p_{m+} \end{pmatrix}$
	2	p_{21}	p_{22}	...	p_{2j}	...	p_{2n}	
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	
	i	p_{i1}	p_{i2}	...	p_{ij}	...	p_{in}	
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	
	m	p_{m1}	p_{m2}	...	p_{mj}	...	p_{mn}	
列の確率 □ (\mathbf{P}_J の対角要素)		p_{+1}	p_{+2}	...	p_{+j}	...	p_{+n}	\uparrow □ 列プロファイルの重心
列の □ 平均ベクトル		$\mathbf{c}_{n \times 1} = (p_{+1}, p_{+2}, \dots, p_{+j}, \dots, p_{+n})^t$						\leftarrow □ 行プロファイルの重心

(行の質量)

(列の質量)

テキスト, 10ページ, 表5

平均ベクトル(重心), 質量(mass)

理論モデル(理論分布)を考える

$$\pi_{ij} = P(I = i, J = j) = P\{I = i \text{ かつ } J = j \text{ となる}\}$$

$$\pi_{i+} = P(I = i), \pi_{+j} = P(J = j) \quad (\text{こう書くことができるから})$$

$$\pi_{ij} = \pi_{i+} \pi_{+j} = P(I = i) P(J = j) \Leftrightarrow \left(\begin{array}{l} \text{項目 } I \text{ と } J \text{ とは} \\ \text{互いに独立というモデル} \end{array} \right)$$

- ① 仮に上のように表せるとしてみる(つまり独立モデルという1つのモデルを仮定). 他のモデルもあり得る.
- ② 「項目 I と項目 J とは関連性がない」としてみようということ.
- ③ これを帰無仮説 H_0 としてカイ二乗統計量を用いて検定.
- ④ (クロス表の)項目間の“独立性の検定”という.

このときのピアソンのカイ二乗統計量

$$\chi_p^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{N(\pi_{ij} - \pi_{i+}\pi_{+j})^2}{\pi_{i+}\pi_{+j}} \quad \left(\begin{array}{l} \text{測定で得た度数から} \\ \text{の実現確率} p \text{を使う, 次ページ参照} \end{array} \right)$$

$$\doteq \sum_{i=1}^m \sum_{j=1}^n \frac{N(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}} = \sum_{i=1}^m \sum_{j=1}^n \frac{\left(f_{ij} - \frac{f_{i+}f_{+j}}{N} \right)^2}{\frac{f_{i+}f_{+j}}{N}}$$

$\left(\begin{array}{l} \text{元の分布} p_{ij} \text{とモデルとした分布} p_{i+}p_{+j} \text{との間} \\ \text{の一種の分布間の距離を測っている} \end{array} \right)$

実現した分布と想定したモデルの一種の分布間の距離になっている。

$$y_{ij} = \frac{f_{ij} - \frac{f_{i+}f_{+j}}{N}}{\sqrt{\frac{f_{i+}f_{+j}}{N}}} \text{を要素とする行列 } \mathbf{Y} = (y_{ij})_{m \times n} \text{の“特異値分解”が対応分析(CA)}$$

(参考)いづれ述べる.

このときの期待度数は, ...

- 期待度数とは独立モデルを理論分布と仮定したときの理論上の期待値のこと.
- 割合の推定, 検定を考えたときに未知の母数を取得の割合から用意することに類似した操作.
- 注: サイコロ投げのように演繹的に特定なモデルを作れるとは限らない.

$$\pi_{ij} \Leftrightarrow p_{ij} = \frac{f_{ij}}{N} \text{ (セル}(i, j)\text{の確率)} \Leftrightarrow f_{ij} = Np_{ij}$$

$$\pi_{i+} = \sum_{j=1}^n \pi_{ij} \Leftrightarrow p_{i+} = \frac{f_{i+}}{N} \text{ (}\pi_{i+}\text{の推定値)}$$

$$\pi_{+j} = \sum_{i=1}^m \pi_{ij} \Leftrightarrow p_{+j} = \frac{f_{+j}}{N} \text{ (}\pi_{+j}\text{の推定値)}$$

（ここで一般に π_{i+} , π_{+j} は未知であるから取得したクロス表の度数から上のよう推定する, これを p_{i+} , p_{+j} で示した）

(つづき)

- このとき, クロス表の各セルの期待度数は次式から得られる.
- 仮説とした「独立」(モデル)の条件から, 積の形となる.
- モデルから推定した“期待度数”を作る, ということだけを知っておくこと.

セル (i, j) の期待度数 e_{ij} は以下で推定する.

$$e_{ij} = Np_{ij} = Np_{i+}p_{+j} = N \times \frac{f_{i+}}{N} \times \frac{f_{+j}}{N} = \frac{f_{i+}f_{+j}}{N}$$



$$N \times [(i \text{の生起確率}) \times (j \text{の生起確率})]$$

ここで, クロス表の度数 f_{ij} から得た推定比率 p_{i+} , p_{+j} を用いて“期待度数”を推定する.

これを覚えておく!!!

$$e_{ij} = \frac{f_{i+}f_{+j}}{N}$$

独立性の検定

$$H_0 : \pi_{ij} = \pi_{i+} \pi_{+j} = P(I = i) P(J = j)$$



[帰無仮説]

(項目 I と J とは互いに独立とするモデル)

- ①この帰無仮説のもとに検定統計量とするピアソンのカイ二乗統計量が“ χ^2 分布に近似すること”を用い検定する.
- ②離散的に分布するピアソンの χ^2 統計量を連続確率分布の χ^2 分布(自由度: $d.f. = (m-1)(n-1)$)で近似すること.
- ③この近似は, 標本の大きさ N がある程度の大きさでよく近似することが分かっている.

自由度 (*d.f.*: degrees of freedom) について

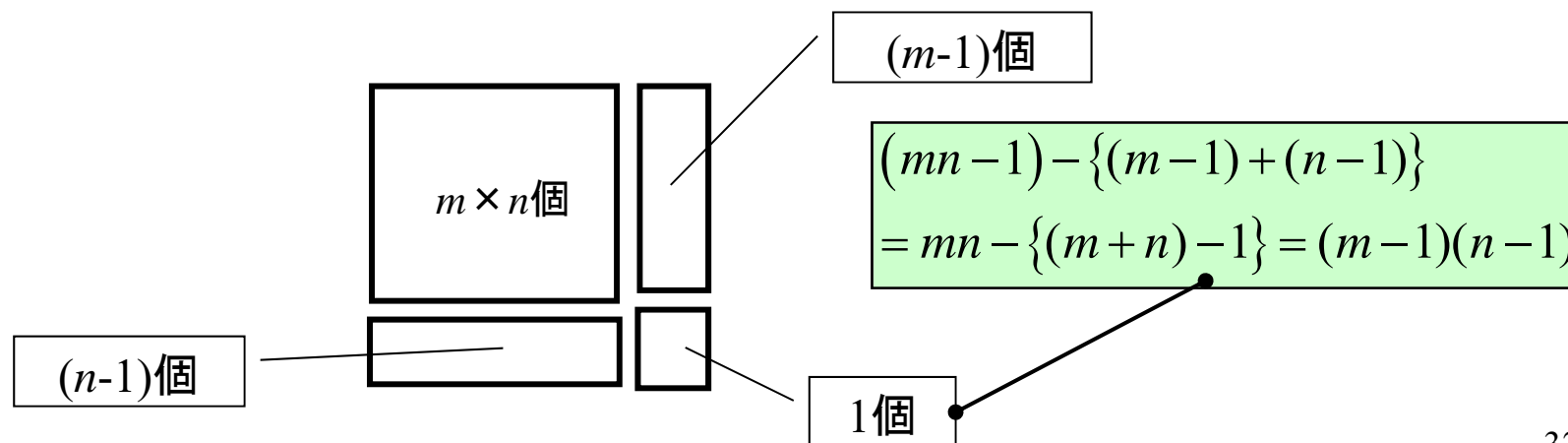
①母比率: p_{ij} はクロス表の行側, 列側の項目 I と J とが互いに無関係 (独立) としたとき, 以下のようになるとした (独立モデル).

$$\pi_{ij} = \pi_{i+} \pi_{+j} \Leftrightarrow P(I=i \text{ かつ } J=j) = P(I=i)P(J=j)$$

②このモデルの仮定のもとに, 手にした実現度数を使うと期待度数は,

$$e_{ij} = Np_{ij} = Np_{i+}p_{+j} = N \times \frac{f_{i+}}{N} \frac{f_{+j}}{N} = \frac{f_{i+}f_{+j}}{N}$$

$$\text{母比率 (の推定値) は, } p_{ij} = p_{i+}p_{+j} = \frac{f_{i+}}{N} \frac{f_{+j}}{N} = \frac{f_{i+}f_{+j}}{N^2}$$



(つづき)

③よって母比率は $(m+n)$ 個の $f_{i+} = \sum_{j=1}^n f_{ij}$, $f_{+j} = \sum_{i=1}^m f_{ij}$ (周辺和の関係)を用いている.

④さらに $\sum_{j=1}^n f_{+j} = \sum_{i=1}^m f_{i+}$ であるから, この中の $(m+n-1)$ 個を決めると残り1個は自ずと決まる. 前ページの図を参照.

⑤以上から, 全体で $m \times n$ 個あるマス(セル)のうち, 上の $(m+n-1)$ 個は拘束されるから(自由でないからそれを除いて),
 $mn - (m+n-1) = (m-1)(n-1)$ となる.
これが“自由度”(d.f.)となる.

この関係を覚えておく!!!

自由度: $d.f. = (m-1)(n-1)$

この自由度(独立に動けるマスの数)を決める経緯として. ピアソンとフィッシャーの有名な論争があった.

当初, ピアソンは自由度: $d.f. = mn - 1$ と考えたらしい.

数値例

- もう一度、前にみた「環境意識調査」データの2つの質問項目を使って確認する.

質問J: あなたは、いま住んでいるまちが気に入っていますか.

1. たいへん気に入っている
2. まあ気に入っている
3. あまり気に入っていない
4. まったく気に入っていない

質問J: あなたの住んでいる地区は、都市としては緑(みどり)が多いと感じますか. それとも少ないと感じますか.

1. かなり多い
2. 多いほうである
3. ふつう
4. 少ないほうだ
5. きわめて少ない

(つづき)

- JMPを用いると、以下の情報が出力される.
- 各セルの「実現度数」(各セルの実際の回答者数): f_{ij}
- 独立モデルとして推定した「期待度数」: e_{ij}
- 「偏差」つまり実現度数と期待度数の差: $f_{ij} - e_{ij}$
- セルごとのカイ二乗の値: $(f_{ij} - e_{ij})^2 / e_{ij}$
- 例: 2つのセルについてカイ二乗値を求めると以下.

$$\text{セル}(1,1) \Rightarrow \frac{(166 - 82.1274)^2}{82.1274} = 85.6548$$

「たいへん気に入っている」かつ
「かなり多い」

$$\text{セル}(4,5) \Rightarrow \frac{(6 - 0.53186)^2}{0.53186} = 56.2188$$

「まったく気に入っていない」かつ
「きわめて少ない」

数値例(JMPの出力例)

Q1(2): まちは気に入っているか_ラベルとQ2(1): 緑が多いか_ラベルの分割表に対する分析

分割表

		Q2(1): 緑が多いか_ラベル					
		かなり多い	多い方である	ふつう	少ない方だ	きわめて少ない	
Q1(2): まちは気に入っているか_ラベル	度数						
	期待値						
	偏差						
	セルのカイ ² 乗						
	たいへん気に入っている	166	239	86	26	7	524
		82.1274	236.689	125.211	61.3936	18.5797	
		83.8726	2.31141	-39.211	-35.394	-11.58	
		85.6548	0.0226	12.2791	20.4045	7.2169	
	まあ気に入っている	131	598	324	146	36	1235
		193.564	557.844	295.105	144.697	43.7898	
		-62.564	40.1557	28.8947	1.30319	-7.7898	
		20.2219	2.8906	2.8292	0.0117	1.3857	
	あまり気に入っていない	6	40	55	51	20	172
		26.9579	77.6917	41.0997	20.1521	6.09866	
		-20.958	-37.692	13.9003	30.8479	13.9013	
		16.2933	18.2859	4.7012	47.2205	31.6868	
	まったく気に入っていない	2	2	0	5	6	15
		2.35098	6.77544	3.58428	1.75745	0.53186	
		-0.351	-4.7754	-3.5843	3.24255	5.46814	
		0.0524	3.3658	3.5843	5.9826	56.2188	
		305	879	465	228	69	1946

各セル内に表示の情報の説明

実際にカイ二乗統計量求めると, ...

$$\chi_p^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(\text{実現度数} - \text{期待度数})^2}{\text{期待度数}} \quad (\text{ここで, } m=4, n=5)$$

$$= \frac{(166 - 82.1274)^2}{82.1274} + \frac{(239 - 236.689)^2}{236.689} + \dots + \frac{(6 - 0.53186)^2}{0.53186} = 340.309$$

$$\chi_{d.f.}^2(0.05) = \chi_{12}^2(0.05) = 21.0261 < \chi_p^2 = 340.309 \quad (\chi^2 \text{分布で近似})$$

$$\left(\begin{array}{l} P(\chi_p^2 \geq \chi_{d.f.}^2) = 0.05 \text{とした, ということ.} \\ \text{ここで } d.f. = (m-1)(n-1) \text{は自由度という.} \end{array} \right)$$

Q1(2): まちは気に入っているか_ラベルとQ2(1): 緑が多いか_ラベルの分割表に対する分析

検定

N	自由度	(-1)*対数尤度	R2乗(U)
1946	12	141.17134	0.0533

(ここでは, 自由度 $d.f. = (4-1)(5-1) = 12$)

検定	カイ2乗	p値(Prob>ChiSq)
尤度比	282.343	<.0001*
Pearson	340.309	<.0001*

数表確認(スライド17p)

これをどう解釈するか

- 検定の結果は、確かに“高度に有意”となった.
- 有意, よって「帰無仮説: 独立である」は棄却された.
- 取り上げた2つの質問項目の間の関係は“関係がない, とはいえないだろう(ありそう)”, その確度がかなり高いようだ. (†)
- しかし, 関係の程度までは分からない. 「関係あり」と断定はできない.

(†) 再確認: 統計的検定に特有の二重否定による解釈(背理法)

(つづき)

- ここである疑問が生じるだろう.

Q1:そもそも, 2つの項目 (I, J) 間に“関連性がありそう”だからこの項目で測定(調査)を行ったのではないか.

Q2:そうであれば「関連がない」という仮説(独立モデル)を前提の議論は, シビアではないのか.

Q3:独立モデル以外にも考えられるのではないか.

- 疑問はいずれも“もつともである”. ではどうするか.
- これを巡ってさまざまな研究が行われてきた.

(つづき)

- “対応分析法”もその1つと考えられる.
- 対応分析では, 2項目の関連性を固有値(成分スコアの分散)という指標で示す.
- その大きさを2つの質問間の関連性を(相関として)知ることができる(特異値で測る). 数量化法Ⅲ類も同じ.
- またピアソンのカイ二乗統計量を固有値の大きさを分解し, 関連の程度を大きさを測っている.
- では“どう考えるのか”, これがこれからのトークの内容.

ここから数ページは, 対応分析法(CA)とは“こんなことを行いたい”という予告・予習として聞いていただく.

予習として, ある関係をチェック

- 所与のクロス表に対応分析法を適用して得られた“固有値の総和”とカイ二乗統計量の関係.

$$\chi_p^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(f_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^m \sum_{j=1}^n \frac{\left(f_{ij} - \frac{f_{i+} f_{+j}}{N}\right)^2}{\frac{f_{i+} f_{+j}}{N}} = \sum_{i=1}^m \sum_{j=1}^n \frac{N(p_{ij} - p_{i+} p_{+j})^2}{p_{i+} p_{+j}}$$

$$\sum_{k=1}^K \lambda_k = \frac{\chi_p^2}{N} = \sum_{i=1}^m \sum_{j=1}^n \frac{(p_{ij} - p_{i+} p_{+j})^2}{p_{i+} p_{+j}} \quad \left(\begin{array}{l} 0 \leq \lambda_k \leq 1, k = 1, 2, \dots, K \\ K = \min\{m, n\} - 1 \end{array} \right)$$

ここで, λ_k は第 k 成分の固有値という.

また, $\sqrt{\lambda_k}$ を特異値という.

$$\phi^2 = \frac{\chi_p^2}{N}$$

平均平方関連係数という.
連関性の測度の1つ.

重要な性質(一部)

◎以下の重要な関係がある. これから何度も登場する.

◎これらを調べること, 述べること, 理解することが, ここからの目標である.

① 対応分析で得た「固有値の総和 = χ^2 / N 」(総変動)となる.

$$\sum_{k=1}^K \lambda_k = \frac{\chi_p^2}{N} (\text{総変動}) \Leftrightarrow \left[\text{固有値の総和} = \frac{\text{ピアソンのカイ二乗統計量}}{N} \right]$$

① 固有値の平方根(特異値という)が成分スコアの相関係数に対応する.

② この相関係数の二乗和がピアソンの χ^2 統計量と関係する.

このあと, 数値例で確かめる. まず記号やその意味をイメージとして覚えておこう.

別の重要な関係

- 以下の“再生公式”(推移公式)という関係がある.
- 実は, R.A. Fisherほかが調べたことでもある.
- ここで, “独立モデル”がどこに現れているか.

$$p_{ij} = p_{i+} p_{+j} \left\{ 1 + \sum_{k=1}^K \frac{1}{\sqrt{\lambda_k}} z_{ik} z_{jk}^* \right\} \quad \left(\begin{array}{l} \lambda_k \text{は固有値(成分スコアの分散)} \\ z_{ik} \text{は行成分スコア, } z_{jk}^* \text{は列成分スコア} \end{array} \right)$$

$$= \underbrace{p_{i+} p_{+j}}_{\text{独立モデルが測る部分}} + p_{i+} p_{+j} \left\{ \sum_{k=1}^K \frac{1}{\sqrt{\lambda_k}} z_{ik} z_{jk}^* \right\} \quad (i \in I, j \in J, K = \min\{m, n\} - 1)$$

独立モデルが測る部分

対応分析が測る部分

$$f_{ij} = \left(\frac{f_{i+} f_{+j}}{N} \right) \left\{ 1 + \sum_{k=1}^K \frac{1}{\sqrt{\lambda_k}} z_{ik} z_{jk}^* \right\} \quad (i \in I, j \in J, K = \min\{m, n\} - 1)$$

2元クロス表: $F = (f_{ij})$

このクロス表に“対応分析法”を適用する(対応分析法で解く).
このデータ表は“多次元データ”であること(行・列の両方向).

形式的にCAを
適用する

$J = \{1, 2, \dots, n\} \Leftrightarrow J = \{1, 2, \dots, 5\}$ 質問 J : 都市としては“緑(みどり)が多い”と感じますか.							
質問 I : いま住んでいるまちが気に入っていますか. $I = \{1, 2, \dots, m\}$ \Updownarrow $I = \{1, 2, \dots, 4\}$	選択肢	1.かなり多い	2.多いほう	3.ふつう	4.少ないほう	5.少ない	行和 (f_{i+})
	1.大変気に入っている	166	239	86	26	7	524
	2.まあ気に入っている	131	598	324	146	36	1,235
	3.あまり気に入っていない	6	40	55	51	20	172
	4.気に入っていない	2	2	0	5	6	15
	列和 (f_{+j})	305	879	465	228	69	1,946 (N)

$$\sum_{i=1}^4 \sum_{j=1}^5 f_{ij} = \sum_{i=1}^4 f_{i+} = \sum_{j=1}^5 f_{+j} = N (=1,946)$$

特異値・固有値，寄与率と累積寄与率

- 特異値・固有値，寄与率，累積寄与率の要約

k	特異値 α_k	固有値 λ_k	寄与率 ν_k	累積寄与率 $\sum_k \nu_k$	累積寄与率 (%)
1	0.35288	0.12452	0.7121	0.7121	72.1
2	0.20959	0.04393	0.2512	0.9633	96.3
3	0.08014	0.00642	0.0367	1.0000	(100)
	固有値の総和 $\sum_{k=1}^3 \lambda_k$	0.17487	—	—	—

$$\alpha_k = \sqrt{\lambda_k}; \sum_{k=1}^3 \lambda_k = 0.17487 \quad (\text{こういう関係にある})$$

ここで“特異値・固有値”とはなにか(何を意味するのか).
なぜ，特異値・固有値はここでは3個なのか.

行・列の選択肢の「成分スコア」を求める

- 成分スコア(行成分スコア, 列成分スコア)を求める.

質問文 $I = \{1, 2, \dots, 4\}$	質問の選択肢	第1成分スコア	第2成分スコア	第3成分スコア
		z_{i1}	z_{i2}	z_{i3}
質問I あなたは、いま住んでいる まちが気に入って いますか.	1.大変気に入っている	-0.4442	0.2027	-0.0353
	2.まあ気に入っている	0.0623	-0.1315	0.0311
	3.あまり気に入っていない	0.7886	0.1907	-0.1698
	4.気に入っていない	1.3458	1.5567	0.6157
質問文 $J = \{1, 2, \dots, 5\}$	質問の選択肢	z_{j1}^*	z_{j2}^*	z_{j3}^*
質問J あなたの住んでいる地区 は、都市としては“緑 (みどり)が多い”と感 じますか.	1.かなり多い	-0.5403	0.3235	-0.0640
	2.多いほう	-0.1118	-0.1055	0.0657
	3.ふつう	0.1545	-0.1506	-0.0613
	4.少ないほう	0.5530	0.0750	-0.1069
	5.少ない	0.9438	0.6805	0.2119

行成分スコア

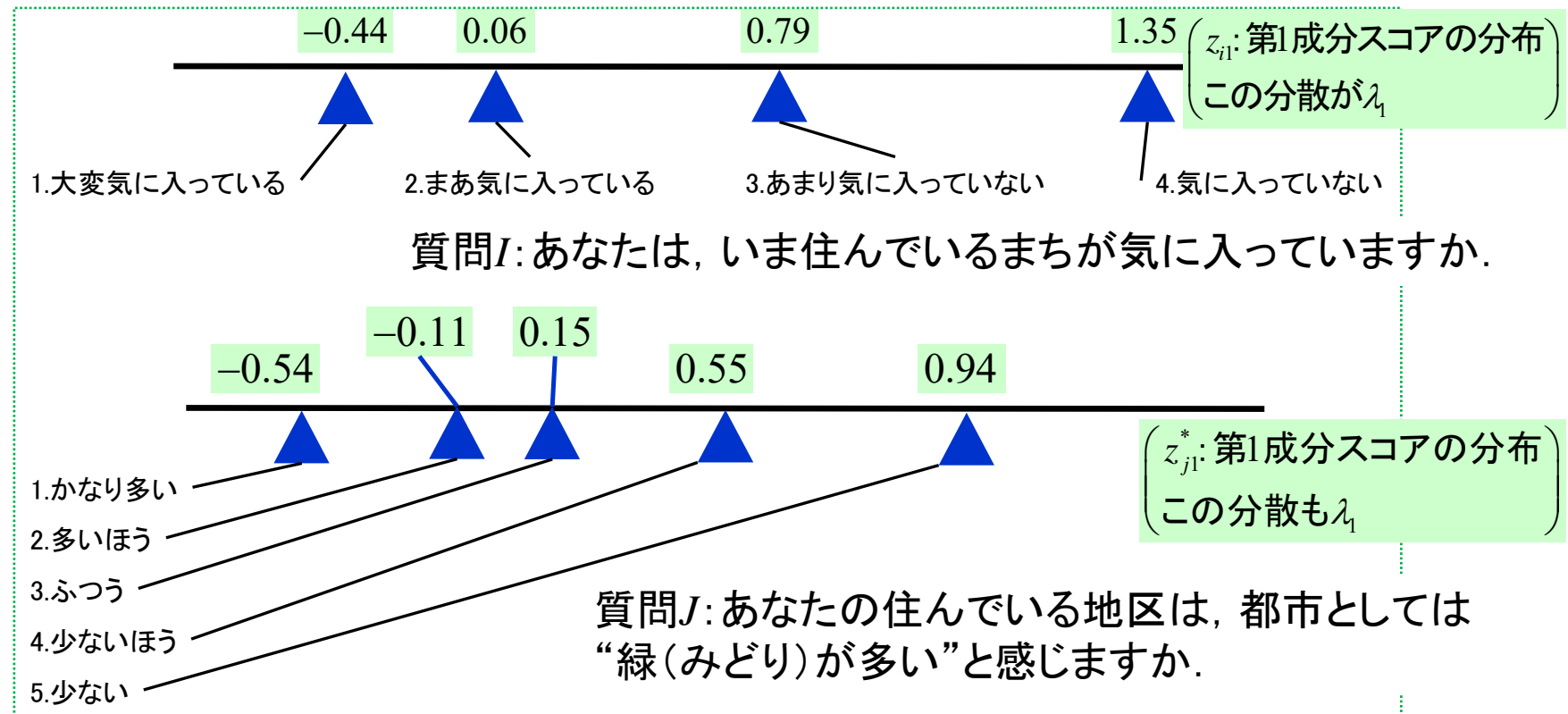
列成分スコア

成分スコア(数量得点)とはどのようにして得られたのか.

成分スコア, 数量得点, 座標といろいろな呼称がある. “coordinate”, “score”の訳.

成分スコアを布置図とする(1次元の布置図)

- “第1成分スコア”に注目, これについて“布置図”を描く.



上のようなことは“何を意味する”のか？

第1成分スコアをもとの選択肢と比べる

- もとの回答選択肢には“名目的なコード”(数値)あるいは選択肢という“標識”(ラベル)が付けられていた.
- 基本は文字変数で“名義尺度”, 質的データである.
- その元のコードとは異なる新たな“数量”(成分スコア)を得たこと.
- ここは(第1成分スコア), 2つの項目とも, “**たまたま並び順は一致**”した.
- いつも, そうなるとは“限らないこと”.
- 並び順が保持されるような尺度が作れるか, 作れたらその元の“標識”を評点(スコア)として使えるかもしれない. ⇔ライカート尺度との関係

(つづき)

- スライド, 前々ページ(図), さらにその前ページ(表)の情報を再確認する.
- これらの数量は“量的データ”として扱えること.
- 四則演算が可能となる. 大きさの比較や別の分析(例: クラスター化)などの処理を容易にする.
- これをさらに“元のクロス表”と比べてみる.
- もとのクロス表には「5行×4列=20セル」あった.
- 各セルの選択肢の組合せとそれへの成分スコアを整理する.

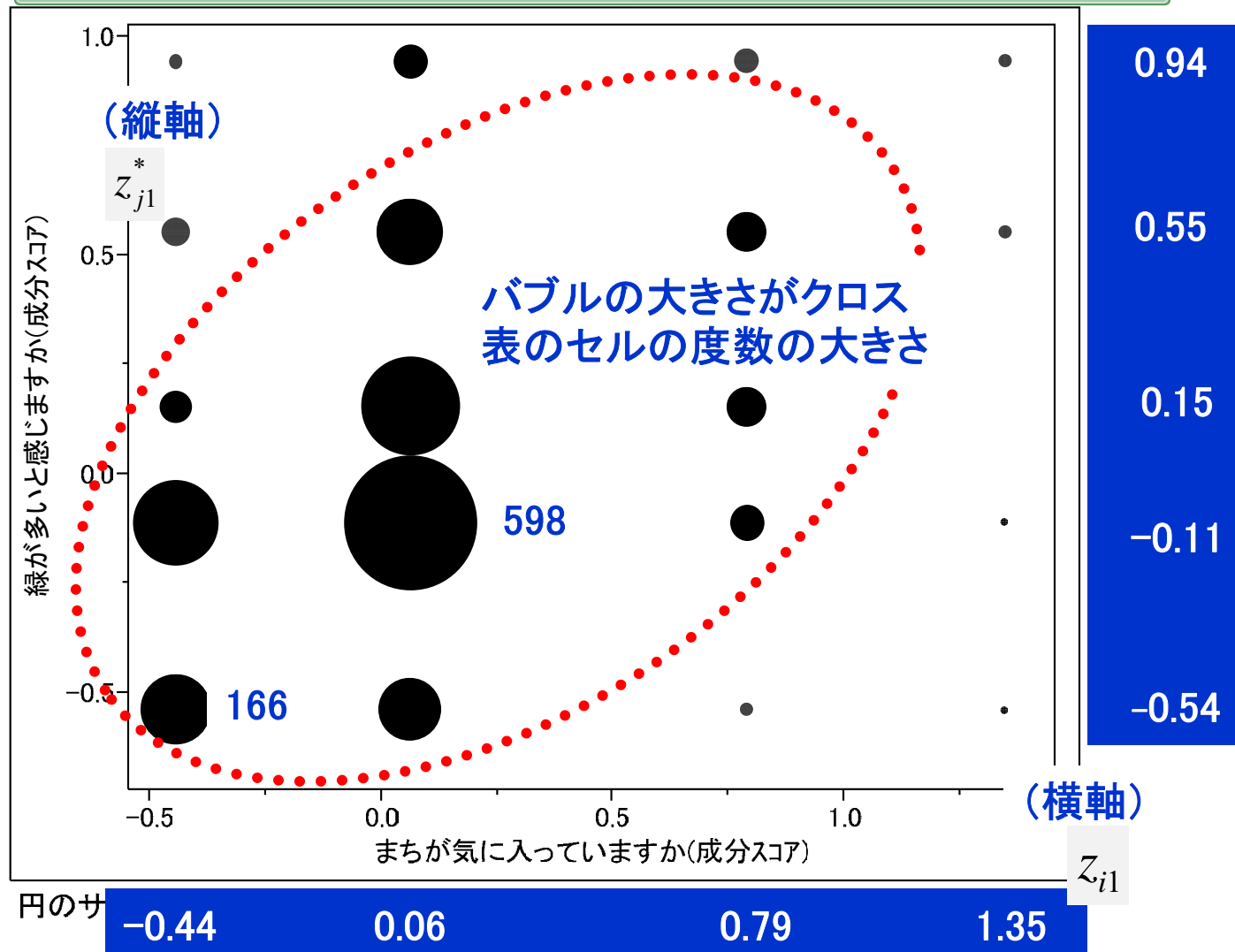
“第1成分スコア”の意味（選択肢との関係）

回答 パター ン	質問I	質問J	回答数（人） [N=1,946]	行の第1成分スコア z_{i1}	列の第1成分スコア z_{j1}^*
1	1.大変気に入っている	1.かなり多い	166	-0.4442	-0.5403
2	1.大変気に入っている	2.多いほう	239	-0.4442	-0.1118
3	1.大変気に入っている	3.ふつう	86	-0.4442	0.1545
4	1.大変気に入っている	4.少ないほう	26	-0.4442	0.5530
5	1.大変気に入っている	5.少ない	7	-0.4442	0.9438
6	2.まあ気に入っている	1.かなり多い	131	0.0623	-0.5403
7	2.まあ気に入っている	2.多いほう	598	0.0623	-0.1118
8	2.まあ気に入っている	3.ふつう	324	0.0623	0.1545
9	2.まあ気に入っている	4.少ないほう	146	0.0623	0.5530
10	2.まあ気に入っている	5.少ない	36	0.0623	0.9438
11	3.あまり気に入っていない	1.かなり多い	6	0.7886	-0.5403
12	3.あまり気に入っていない	2.多いほう	40	0.7886	-0.1118
13	3.あまり気に入っていない	3.ふつう	55	0.7886	0.1545
14	3.あまり気に入っていない	4.少ないほう	51	0.7886	0.5530
15	3.あまり気に入っていない	5.少ない	20	0.7886	0.9438
16	4.気に入っていない	1.かなり多い	2	1.3458	-0.5403
17	4.気に入っていない	2.多いほう	2	1.3458	-0.1118
18	4.気に入っていない	3.ふつう	0	—	—
19	4.気に入っていない	4.少ないほう	5	1.3458	0.5530
20	4.気に入っていない	5.少ない	6	1.3458	0.9438

もとのクロス表の各セル内の度数の書き替え（5×4=20セルあったことを想起）

2項目の“第1成分スコア”を布置図とする

まちが気に入っていますか(成分スコア)と緑が多いと感じますか(成分スコア)のバブルプロット サイズ:回答数



次のような関係にある

第 1 成分の固有値 [第 1 成分スコアの分散 (慣性)]

$$\lambda_1 = 0.12452$$

第 1 成分スコアの特異値 [第 1 固有値の正の平方根]

$$\alpha_1 = \sqrt{\lambda_1} = \sqrt{0.12452} = 0.35288$$

行と列の第 1 成分スコアの相関係数 [特異値]

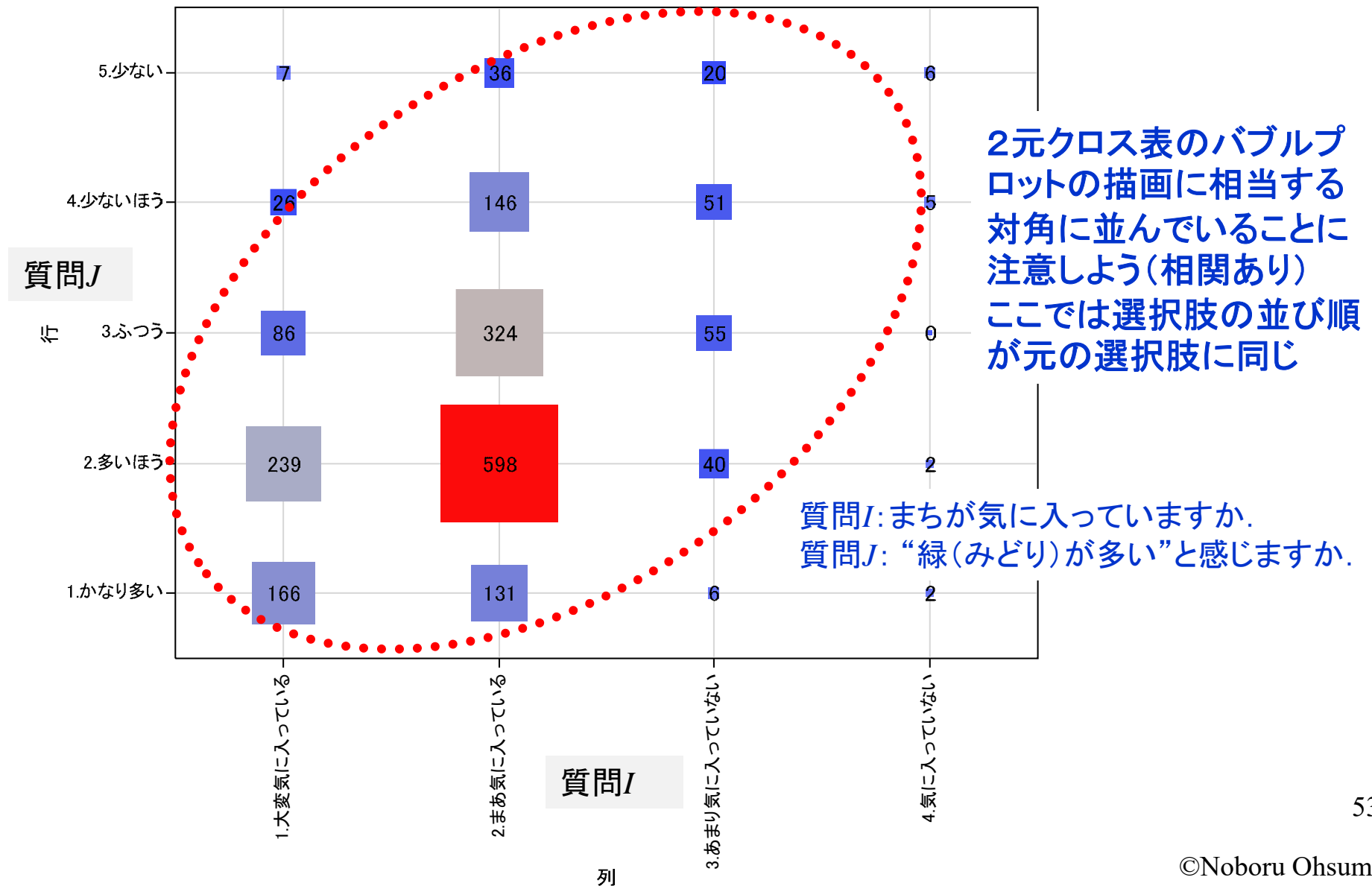
$$\text{Cor}(z_{i1}, z_{j1}^*) = 0.35288$$

ここで,

行の第 1 成分スコア : z_{i1} ($i = 1, 2, \dots, 4$),

列の第 1 成分スコア : z_{j1}^* ($j = 1, 2, \dots, 5$)

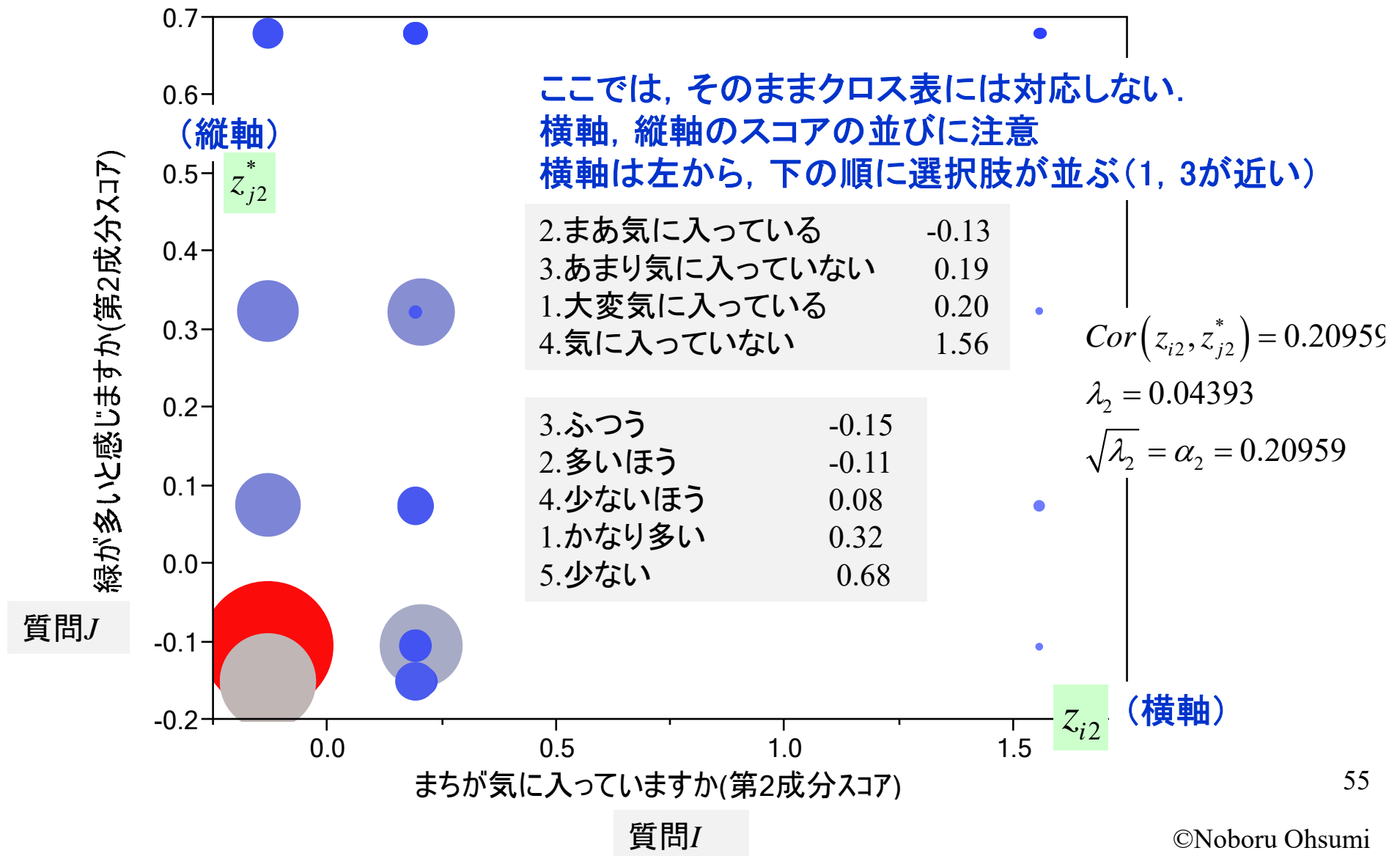
第1成分スコアの並べ替え(クロス表の視覚化)



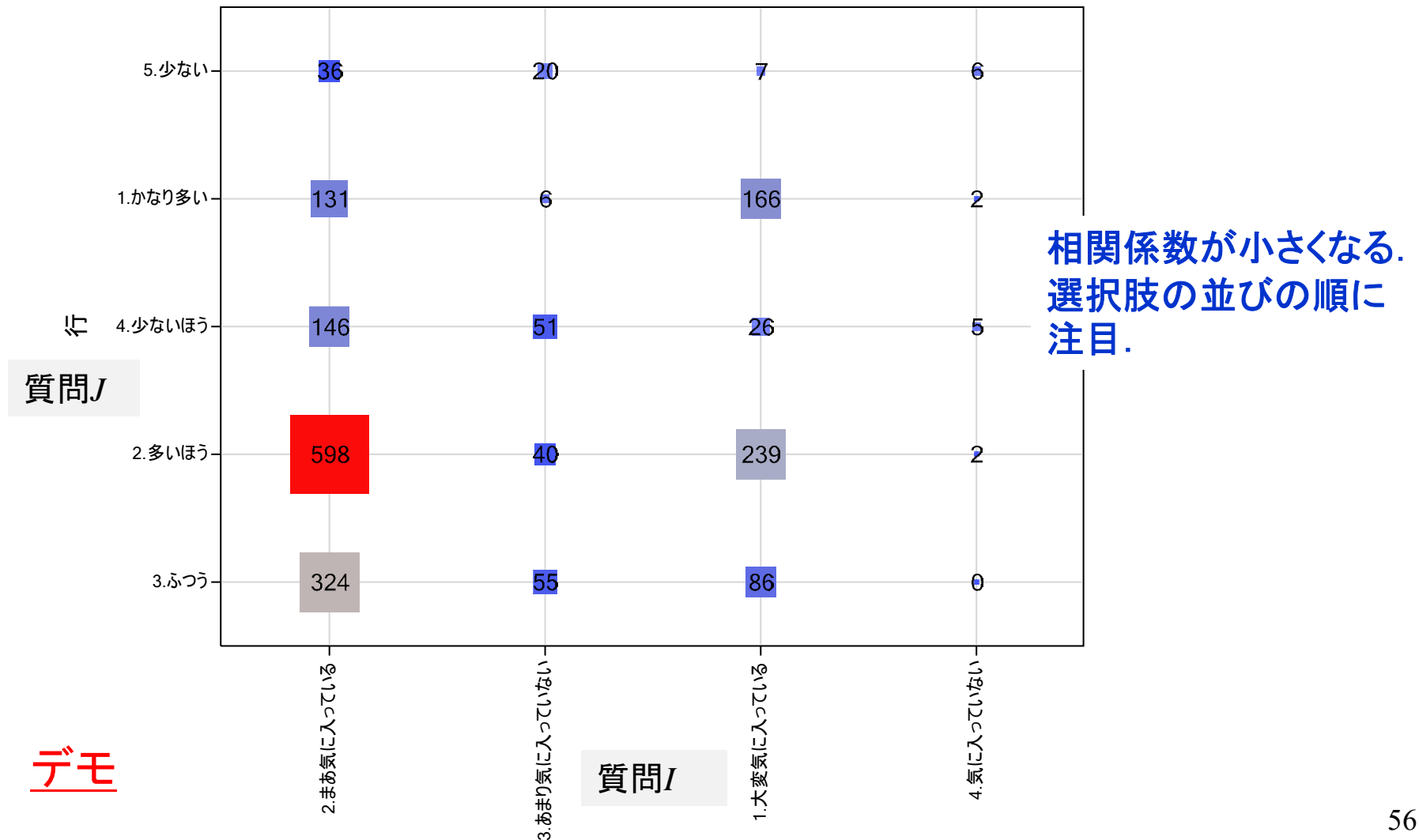
“第2成分スコア”の意味（選択肢との関係）

回答 パターン	質問I	質問J	回答数（人） [N=1,946]	行の第1成分スコア z_{i2}	列の第1成分スコア z_{j2}^*
1	1.大変気に入っている	1.かなり多い	166	0.2027	0.3235
2	1.大変気に入っている	2.多いほう	239	0.2027	-0.1055
3	1.大変気に入っている	3.ふつう	86	0.2027	-0.1506
4	1.大変気に入っている	4.少ないほう	26	0.2027	0.0750
5	1.大変気に入っている	5.少ない	7	0.2027	0.6805
6	2.まあ気に入っている	1.かなり多い	131	-0.1315	0.3235
7	2.まあ気に入っている	2.多いほう	598	-0.1315	-0.1055
8	2.まあ気に入っている	3.ふつう	324	-0.1315	-0.1506
9	2.まあ気に入っている	4.少ないほう	146	-0.1315	0.0750
10	2.まあ気に入っている	5.少ない	36	-0.1315	0.6805
11	3.あまり気に入っていない	1.かなり多い	6	0.1907	0.3235
12	3.あまり気に入っていない	2.多いほう	40	0.1907	-0.1055
13	3.あまり気に入っていない	3.ふつう	55	0.1907	-0.1506
14	3.あまり気に入っていない	4.少ないほう	51	0.1907	0.0750
15	3.あまり気に入っていない	5.少ない	20	0.1907	0.6805
16	4.気に入っていない	1.かなり多い	2	1.5567	0.3235
17	4.気に入っていない	2.多いほう	2	1.5567	-0.1055
18	4.気に入っていない	3.ふつう	0	—	—
19	4.気に入っていない	4.少ないほう	5	1.5567	0.0750
20	4.気に入っていない	5.少ない	6	1.5567	0.6805

“第2成分スコア”の並べ替え



第2成分スコアのクロス表(イメージ)は, ...



JMRAマーケティング・リサーチ講座

質的データのマイニング のための対応分析法

【復習と補足】

トーク内容の復習と要約

トイ・データによる探査

とくに総変動とプロフィールの関係

大隅 昇

ohsumi@ss.ij4u.or.jp

<http://wordminer.org/>

Copyright by Noboru Ohsumi

ここまでのトークの要約

- いろいろと述べた(ようにみえる). 実は同じ内容を“見方を変えて”繰り返し説明した.
- 復習として, ここまでに述べたことを圧縮・要約する.
- 対応分析にはさまざまな変形がある. ここではもっとも基本的な(対称型の)対応分析について説明した.
- とくに以下の言葉の意味
 - ①ピアソンのカイ二乗統計量と“総変動”(全慣性)の関係
 - ②総変動とはなにか
 - ③なぜプロファイル, ストレッチ・プロファイルか, その役割
 - ④プロファイルの重心座標系と総変動の関係
 - ⑤成分スコア(合成指標)の意味, その統計量(平均, 分散)
 - ⑥主な寄与度(絶対寄与度, 相対寄与度)

対応分析法は何を行うのか？

- 基本は“2元データ”表を“多次元データ”としてどう考えるか, にある.
- いま, “寸法が $m \times n$ の2元データ表”を想定する.
- たとえば, 2元クロス表はその1つである.
- いくつかの要件を満たせば対応分析の適用対象となる.

$$\left(\begin{array}{l} \mathbf{F} = (f_{ij})(i \in I, j \in J) \\ I = \{1, 2, \dots, i, \dots, m\} \\ J = \{1, 2, \dots, j, \dots, n\} \end{array} \right) \Leftrightarrow$$

		項 目 J						
	選択肢	1	2	...	j	...	n	行和
項 目 I	1	f_{11}	f_{12}	...	f_{1j}	...	f_{1n}	f_{1+}
	2	f_{21}	f_{22}	...	f_{2j}	...	f_{2n}	f_{2+}
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	i	f_{i1}	f_{i2}	...	f_{ij}	...	f_{in}	f_{i+}
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	m	f_{m1}	f_{m2}	...	f_{mj}	...	f_{mn}	f_{m+}
	列和	f_{+1}	f_{+2}	...	f_{+j}	...	f_{+n}	f_{++}

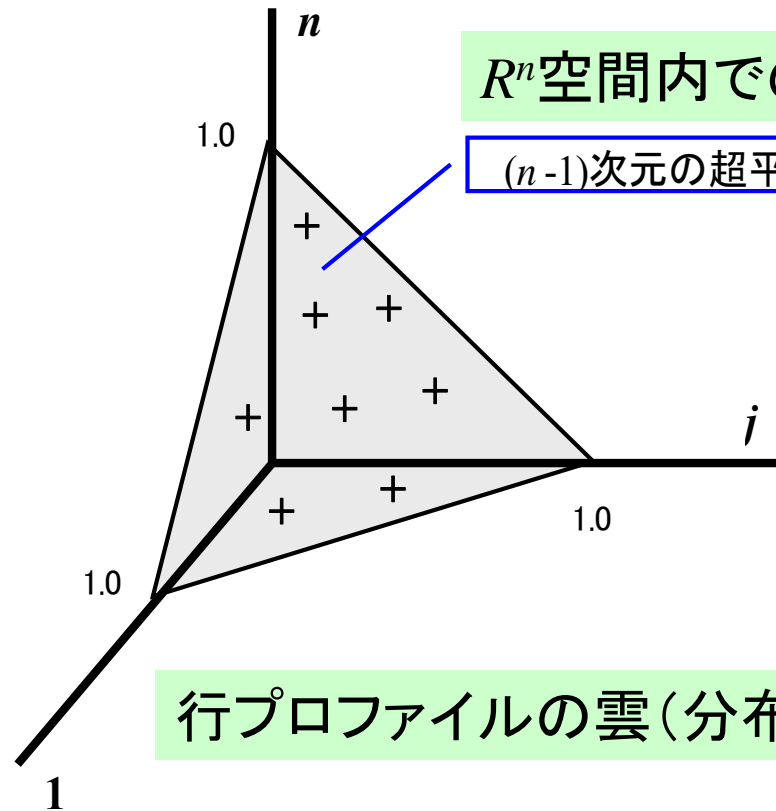
基本のデータはなにか？

- 2元データ表の“(行と列の)プロフィール”を考える.
- プロファイルとは行または列の“相対度数”(実現確率)のこと.
- このプロフィールの分布(“雲”)を考える.
- これが“多次元空間”(重心座標系)に布置する点と考える.
- さらにストレッチ・プロフィールの布置空間を考える.
- この空間内で点の示す分布の“総変動”を考える.
- “総変動”つまり点のチラバリとその大きさでデータ構造の特徴を調べる. これは対応分析にかぎらず統計手法の常套手段.

さらに、対応分析では、…

- 表の行側の m 個の点（項目 I の m 個の選択肢： $i \in I$ ）が，“ n 次元の空間（ R^n ）に分布する”と考えるとき.
- 表の列側の n 個の点（項目 J の n 個の選択肢： $j \in J$ ）が，“ m 次元の空間（ R^m ）に分布する”と考えるとき.
- 実際には、（数理的制約から） $K = \min\{m, n\} - 1$ ，つまりこの次元空間内に分布すること.
- どちらから考えても同じ結果となること（ここでいう対称型の対応分析の場合）.
- これを図で再確認する．またトイ・データで確認する．

プロファイルの空間(重心座標系)



R^n 空間内での分析

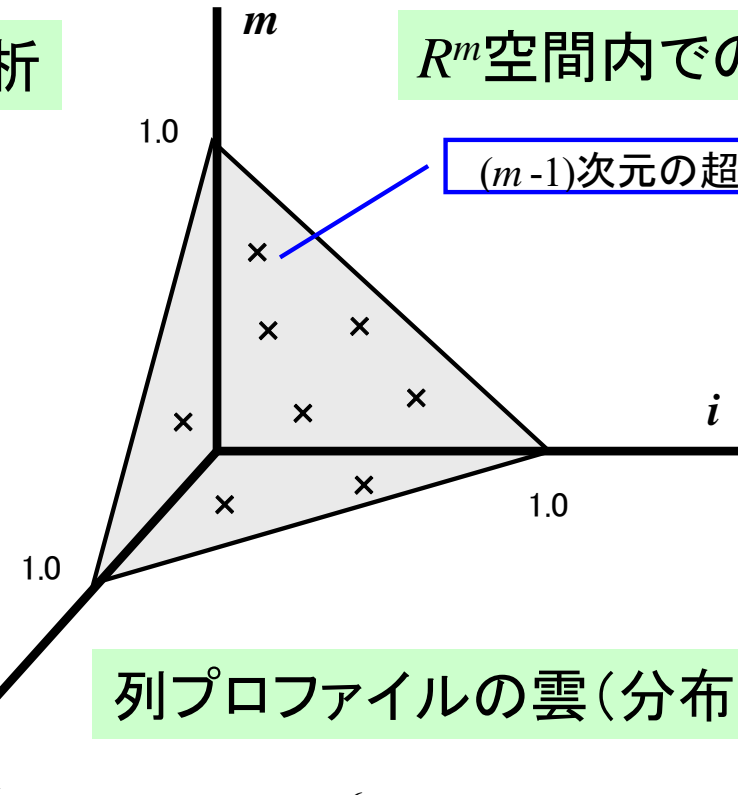
$(n-1)$ 次元の超平面

行プロファイルの雲(分布)

$$\mathbf{N}_I = \mathbf{P}_I^{-1} \mathbf{P}_{IJ} = \left\{ q_{ij} = \frac{p_{ij}}{p_{i+}} = \frac{f_{ij}}{f_{i+}} \mid i \in I, j \in J \right\}$$

$m \times n$ $m \times m$ $m \times n$

$(n-1)$ 次元空間に分布する m 個の点



R^m 空間内での分析

$(m-1)$ 次元の超平面

列プロファイルの雲(分布)

$$\mathbf{N}_J = \mathbf{P}_J^{-1} \mathbf{P}_{JI} = \left\{ q_{ij}^* = \frac{p_{ij}}{p_{+j}} = \frac{f_{ij}}{f_{+j}} \mid i \in I, j \in J \right\}$$

$n \times m$ $n \times n$ $n \times m$

$(m-1)$ 次元空間に分布する n 個の点

いくつかのトイ・データを比べる

- 重心座標系での点(プロフィール)の分布とその総変動の関係を観察する, ミニチュア・データを作る.
- 5種類の2元データ表を用意し比べる.
- データ表の寸法はすべて, $m=5, n=3$ とする.
- つまり, 2次元平面内($n-1=2$)の三角座標図で“全情報”(総変動の全情報)を視認できる.
- 5つのデータ表の“総変動”と分布の様子がどう異なるかを調べる.
- とくに, 「サンプル⑤」の点の分布の特徴に注目する.

サンプル・データの2元データ表

- 「サンプル①」～「サンプル④」の4つのデータ表.
- 列和を100とした. 行プロファイルの重心を同じとした.
- つまり, 行の平均(重心)であり列の質量は同じとする.
- よって, ストレッチ・プロファイルも正三角形図となる.

サンプル①

標識	X1	X2	X3	行和
A	22	20	21	63
B	18	22	17	57
C	20	16	20	56
D	16	18	24	58
E	24	24	18	66
列和	100	100	100	300

サンプル③

標識	V1	V2	V3	行和
A	32	8	6	46
B	8	38	8	54
C	38	12	4	54
D	12	18	70	100
E	10	24	12	46
列和	100	100	100	300

サンプル②

標識	Y1	Y2	Y3	行和
A	24	16	14	54
B	10	24	22	56
C	28	16	18	62
D	16	18	36	70
E	22	26	10	58
列和	100	100	100	300

サンプル④

標識	W1	W2	W3	行和
A	40	2	3	45
B	2	45	2	49
C	46	4	5	55
D	4	3	85	92
E	8	46	5	59
列和	100	100	100	300

(つづき)

- これは他のサンプルデータと比べ、どう違うか. あるいはなにが同じか, 異なるか.

サンプル⑤

標識	Z1	Z2	Z3	行和
A	25	6	2	33
B	35	20	3	58
C	30	28	30	88
D	6	38	55	99
E	4	8	10	22
列和	100	100	100	300

各データ表の「総変動」とは何か？

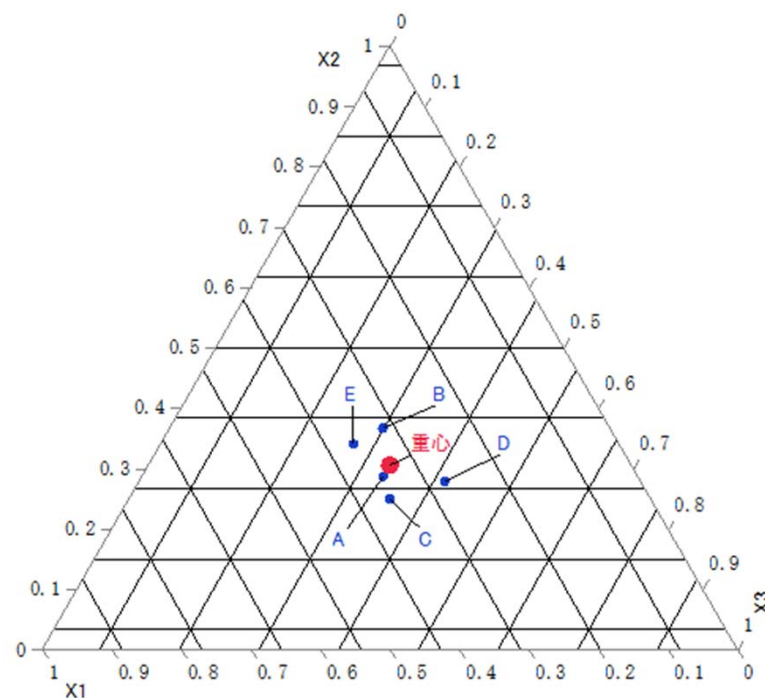
- 各データ表のプロファイルの関係を“三角座標図”で描いてみる(2次元空間に入るから).
- 5つのデータ表から求めた“総変動”(全慣性)を要約する.
下の表. 総変動(全慣性) = 固有値の総和である.
- ここで, 各布置図内の点(行プロファイル)の分布のちらばりの様子, つまり“総変動”の大きさを比べる.
- サンプル①<サンプル②<サンプル③<サンプル④の順に大きくなること. 図で確認しよう.

データ名	総変動(全慣性)
サンプル①	0.01429
サンプル②	0.10275
サンプル③	0.54756
サンプル④	1.32730
サンプル⑤	0.31374

- 行の各点(英文字標識)と重心との距離(ここは**カイ二乗距離**)と点のチラバリ方の程度に注目しよう.
- これと総変動との関係はどう読む？
- 点のチラバリが大きくなると(拡がると), 総変動は大きくなるということ.
- 重心と行の各点(英文字標識)との(質量の重み付き)平方距離の総和が総変動の大きさ(点のチラバリの大きさ)に合わせ変化する.
- また, 各行の標識(英文字)と列(英文字)である頂点の関係に注目する. 近いか遠いか.

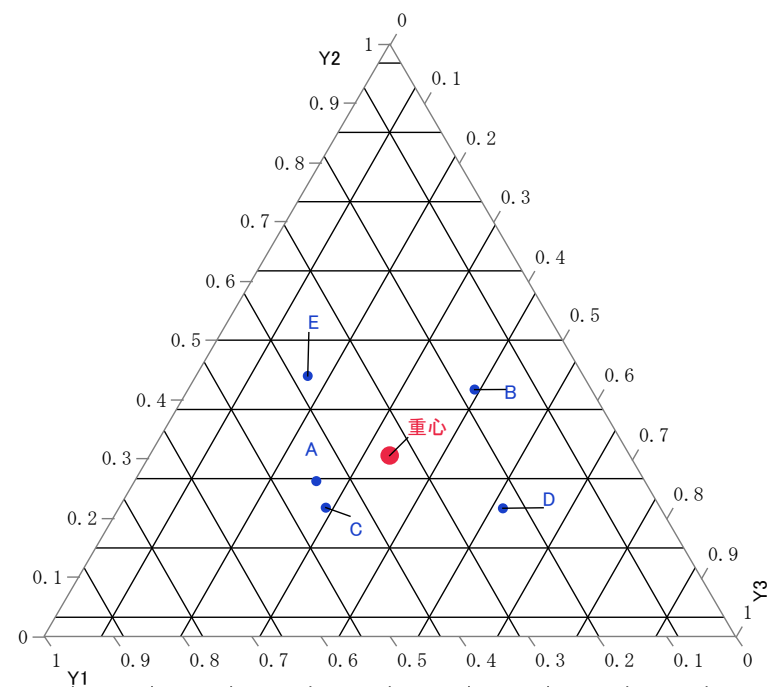
データの三角図を観察

- サンプル①とサンプル②を描画. 点の分布に注目.
- その対応分析で得た固有値他の情報を並記.



結果						
次元	特異値	固有値	割合(%)	割合のプロット	累積(%)	累積のプロット
1	0.10418	0.01085	75.9		75.9	
2	0.05863	0.00344	24.1		100	
固有値の合計 = 0.0142917377037159						

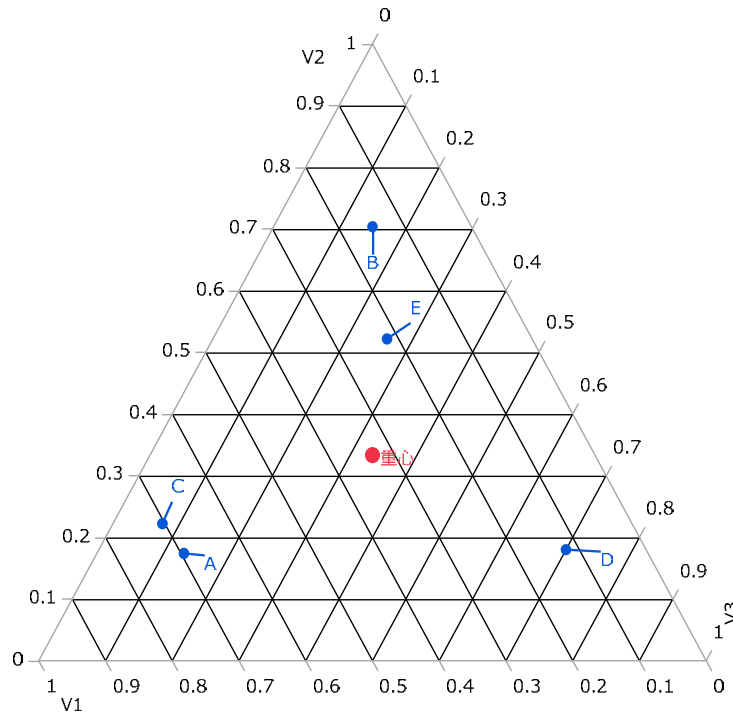
サンプル①のとき



結果						
次元	特異値	固有値	割合(%)	割合のプロット	累積(%)	累積のプロット
1	0.26818	0.07192	70.0		70.0	
2	0.17559	0.03083	30.0		100	
固有値の合計 = 0.102754606823572						

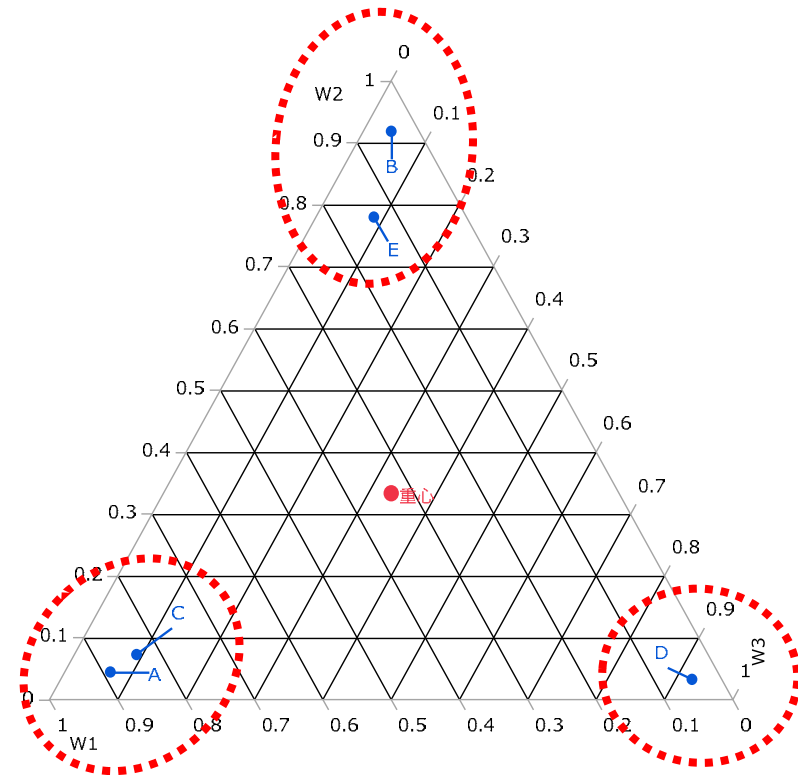
サンプル②のとき

- サンプル③とサンプル④を描画. 点のチラバリがさらに大きくなる. 各図の違いは何か？



結果					
次元	特異値	固有値	割合(%)	割合のプロット	累積(%)
1	0.59104	0.34933	63.8		63.8
2	0.44523	0.19823	36.2		100
固有値の合計 = 0.547556843800322					

サンプル③のとき



結果					
次元	特異値	固有値	割合(%)	割合のプロット	累積(%)
1	0.83733	0.70112	52.8		52.8
2	0.79131	0.62618	47.2		100
固有値の合計 = 1.32729651363013					

サンプル④のとき

サンプル②とサンプル④の同時布置図

- 対応分析を行って得た成分スコアの“同時布置図”を描画する.
- 2組のデータ表の固有値と寄与率を観察しよう. どうなっているか.
- 同時布置図内の「行の標識」(A~E)と列の標識(英文字X1, X2など)の関係は, 同時布置図にどう射影されたか.
- また, 行と列の標識の関係は, 三角図の上と同時布置図のうえでどう表現されているか.
- たとえば, 前ページのサンプル④はどうか(赤点線囲みを観察).

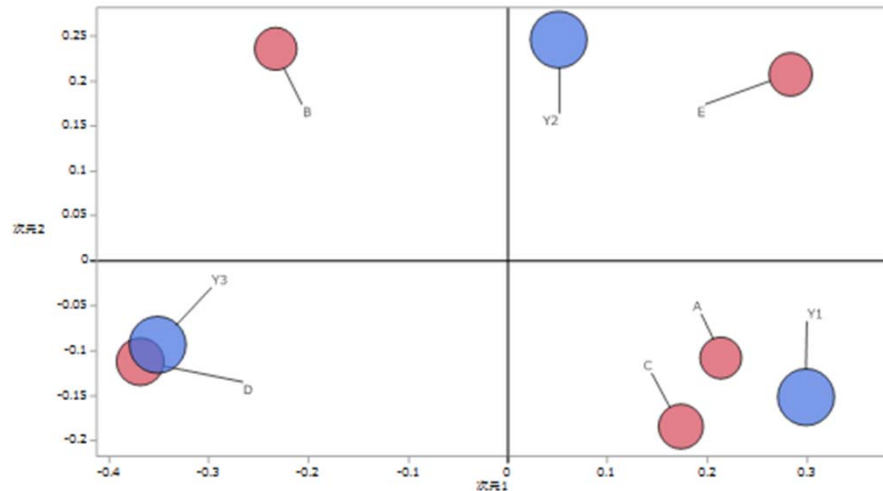
(つづき)

- 三角図の各頂点(列側標識)と図内の行側標識の関係を観察すること.
- 重心から各点(レストラン)までの平方カイ二乗距離を考える.
- この総和が“総変動”となる.
- 一方, 成分スコア (z_{i1}, z_{i2} と z_{j1}^*, z_{j2}^*) の分散(固有値)の和も“総変動”である(そうなるよう成分軸を決めた).

$$\begin{aligned}\text{総変動(全慣性)}: \phi^2 &= \frac{\chi_p^2}{N} \\ &= \sum_{i=1}^m (\text{クロス表の第}i\text{行の質量}) \times \left[\begin{array}{l} \text{第}i\text{行プロフィールと行の} \\ \text{平均プロフィール(重心)との}\chi^2\text{ 距離} \end{array} \right] \\ &= \sum_{j=1}^n (\text{クロス表の第}j\text{行の質量}) \times \left[\begin{array}{l} \text{第}j\text{行プロフィールと} \\ \text{列の平均プロフィール(重心)との}\chi^2\text{ 距離} \end{array} \right]\end{aligned}$$

$$\text{総変動(全慣性)}: \frac{\chi_p^2}{N} = \sum_{k=1}^K \lambda_k \quad (\text{ここで, } K = \min\{m, n\} - 1)$$

サンプル②と④の同時布置図と固有値ほか

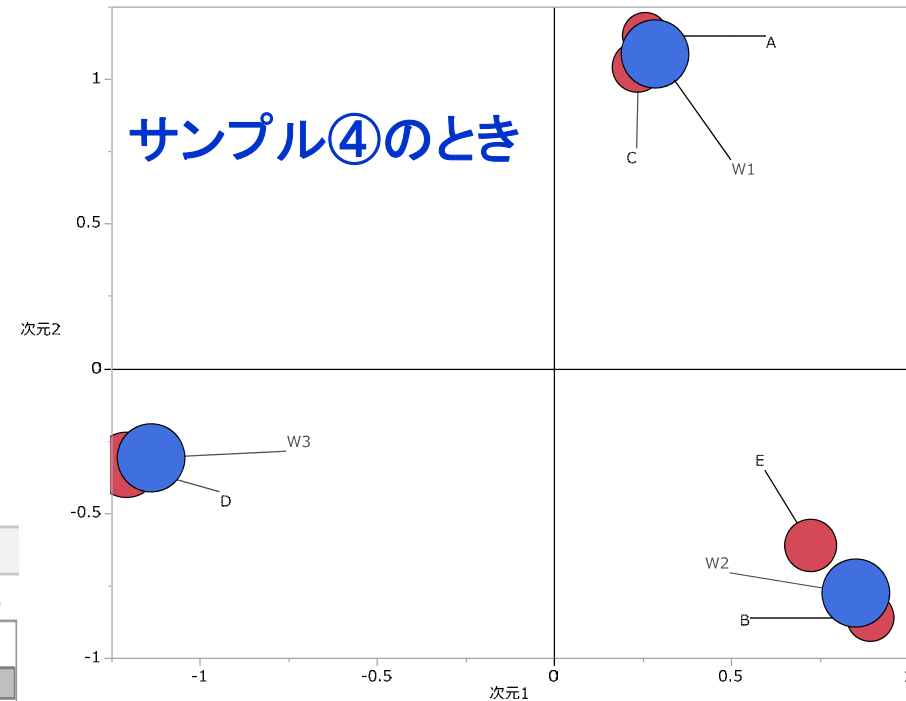


結果

次元	固有値	固有値	割合(%)	割合のプロット	累積(%)	累積のプロット
1	0.26818	0.07192	70.0		70.0	
2	0.17559	0.03083	30.0		100	

固有値の合計 = 0.102754606823572

サンプル②のとき



結果

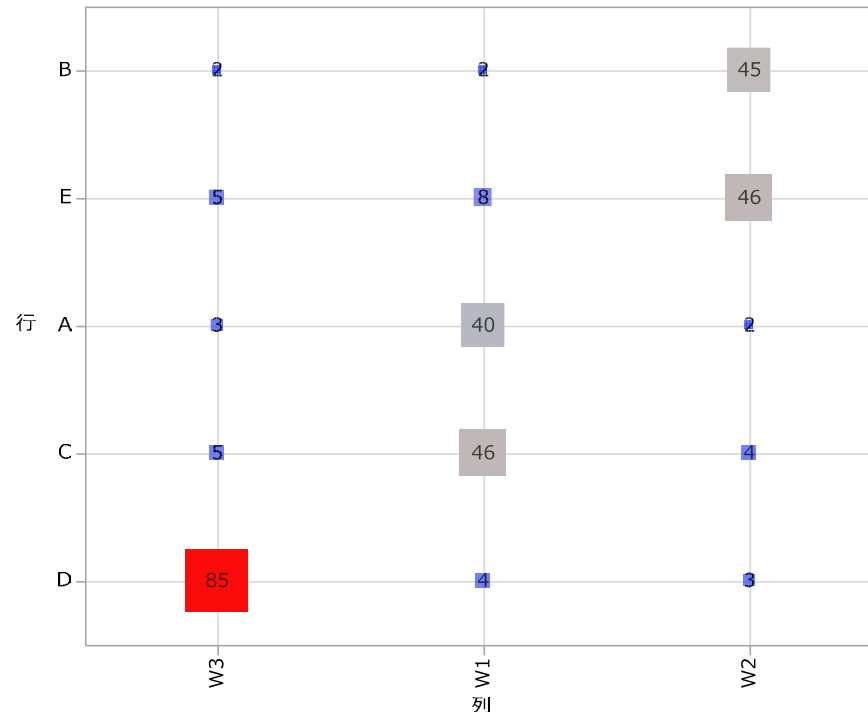
次元	固有値	固有値	割合(%)	割合のプロット	累積(%)	累積のプロット
1	0.83733	0.70112	52.8		52.8	
2	0.79131	0.62618	47.2		100	

固有値の合計 = 1.32729651363013

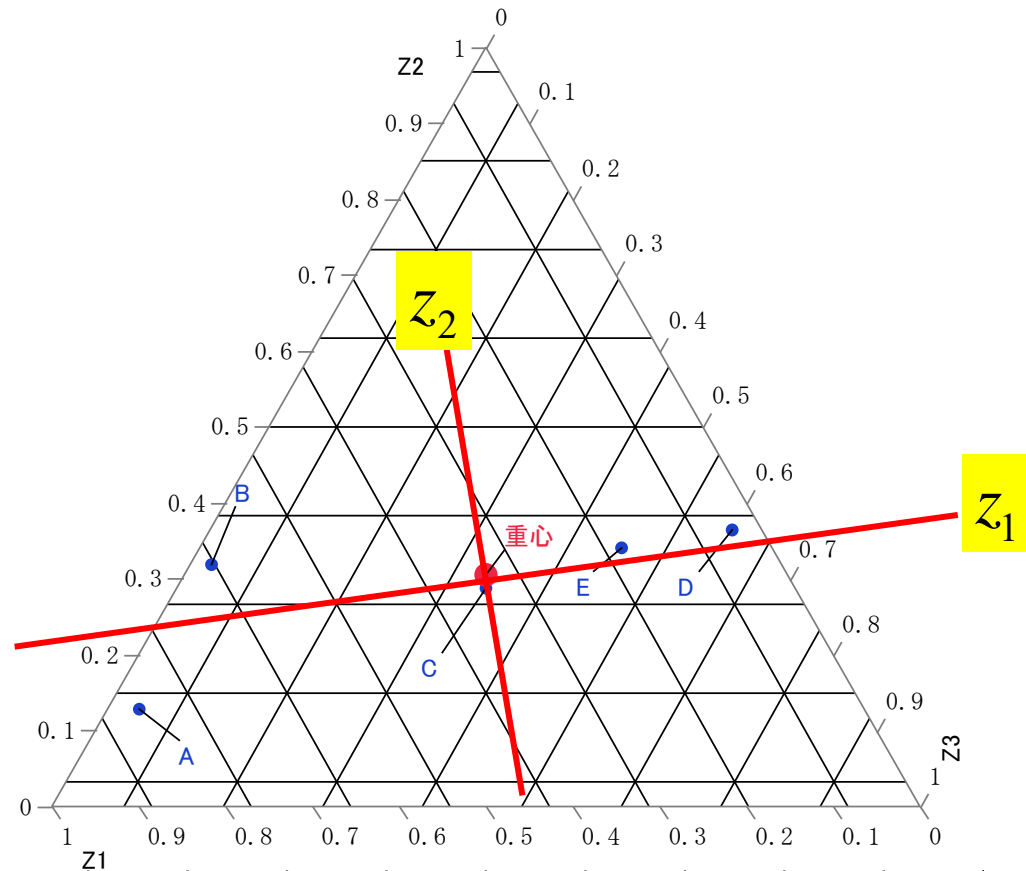
- 固有値(成分スコアの分散)と図の縦横比の關係に注目.
- 行要素と列要素の布置の關係, 元の三角図と比較する.

サンプル④のバブルプロットを描く

- このデータのバブルプロット(クロス表のイメージ)を描く.
- 相関のあること, 行・列の標識の並び順の入れ換えを確認する.
- これを第1成分スコアで散布図とするとその相関係数が特異値である(値を確認しよう).



では、サンプル⑤はどうなるか



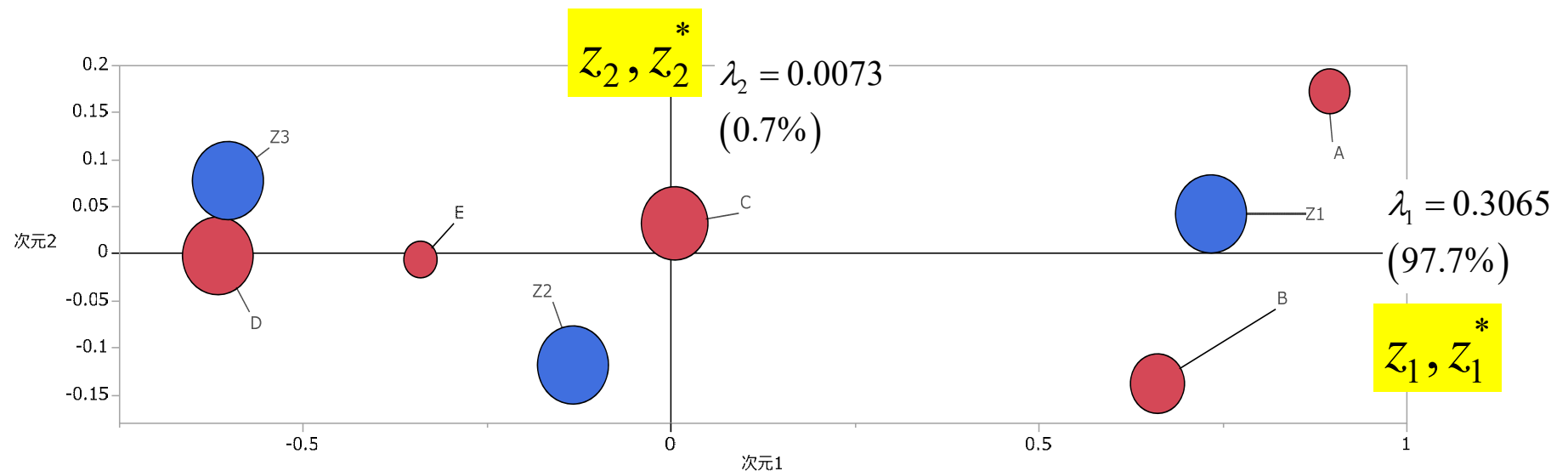
- 点の分布に“ある構造”を埋め込んでみた.
- ここで点の分布はどうなっているか.
- “重心”(ここは行の平均)のまわりで軸を“回転”してみよう. (z_1, z_2)を作る.
- その軸からみた布置と次ページの布置図を比べる.
- 点の分布を“視点を変えて”みること.

結果

次元	特異値	固有値	割合(%)	割合のプロット	累積(%)	累積のプロット
1	0.5536	0.30647	97.7		97.7	
2	0.08526	0.00727	2.3		100	

固有値の合計 = 0.313744339951236

- ここは“2次元”であるから、全情報（総変動）が再現される.
- 通常は多次元空間であるから、分散の大きい成分軸から“探す”. これはソフトが行う.
- かりに“全成分軸”を探せば、それを(固有値を)加えた和が“総変動”となる. K 個の成分 ($K = \min\{m, n\} - 1$) がある.



- 縮約がうまく機能すれば(≡点の布置構造に特徴がある), 分散の大きいはじめのほうのいくつかの成分で説明ができる(次元の縮約化). ここでは第1成分で約97%となる.
- 少数成分となっても, 各行(列)成分スコアには, 列(行)側の情報が加重和として反映されている(多次元情報が反映).

$$z_{ik} = \sum_{j=1}^n l_{jk} x_{ij} = \sum_{j=1}^n \left(\frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} \right) l_{jk} \quad (i \in I; k = 1, 2, \dots, K)$$

$$z_{jk}^* = \sum_{i=1}^m u_{ik} x_{ij}^* = \sum_{i=1}^m \left(\frac{p_{ij}}{p_{+j} \sqrt{p_{i+}}} \right) u_{ik} \quad (j \in J; k = 1, 2, \dots, K)$$

選択肢のストレッチ・プロファイルの加重和
固有ベクトルあるいは特異ベクトルの要素が加重の“係数”

- ここで、“ある構造”としたが、これが行と列との“関連性”に関係する.
- つまり、行の標識と列の標識に付与した成分スコア(数量)の間に相関があれば(正確には線形性の相関),それを意味ある成分軸として“抽出できる”はずである.
- 言い換えると,(堂々巡りだが)“うまい構造を引き出せる”データ収集が重要ということ.
- 調査であれば,うまい質問項目の設計であり適切な構成概念,測定であれば,現象を説明出来そうな複数の適切な変量を選ぶこと(多変量特性).

例:グリッド その1 その2

「ストレッチ・プロファイル」と「カイ二乗距離」

- なぜ, ストレッチ・プロファイルとそれを用いたカイ二乗距離を考えるのか.
- 答え(の一部)は, すでに述べてある. 数理的な証明が必要だが, これは略した(テキストに一部あり).
 - ① “双対性” が成立するため
 - ② “分布の同等性” がなり立つため
 - ③ ストレッチ・プロファイル間距離を“(平方)カイ二乗距離”とすることで, 成分スコア間の距離は“(平方)ユークリッド距離”となること. クラスタ化に関係.

以下の重要な性質

- 成分スコア間の“双対性”と“推移公式”(下の式のこと).
- 双対性は同時布置図の解釈や“追加処理”で重要な役割をはたす.
- “分布の同等性”がある. これは簡単な例を示した. プロファイルが同じ行あるいは列は併合しても対応分析の結果は変わらない.

$$\underline{z_{ik}} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^n \left(\frac{p_{ij}}{p_{i+}} \right) \underline{z_{jk}^*} \quad (i \in I; k = 1, 2, \dots, K)$$

(“たすきがけ”になっている)

$$\underline{z_{jk}^*} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^m \left(\frac{p_{ij}}{p_{+j}} \right) \underline{z_{ik}} \quad (j \in J; k = 1, 2, \dots, K)$$

「寄与度」について

- 寄与度 (contribution) と名付けられた指標は多数ある.
- 繰り返すが, 限られた次元の布置図, 同時布置図から, 主観的観察でなにかを判断してはいけない(誤用のおそれ).
- 基本は“絶対寄与度”と“相対寄与度”(平方相関)である.
- 絶対寄与度: “ある成分 k の軸の解釈”に用いる指標. ある点(選択肢)の各成分への寄与の程度.
- 相対寄与度: ある点(選択肢)が, 各成分軸により, どの程度“近似されるか”(軸の説明力があるか)を示す指標.
- この程度に覚えておく. 例で確かめること(演習).
- 多次元情報を観察するために必要であること.

(つづき)

- 布置図, 同時布置図でみている情報は多次元情報の“一部”にすぎない.
- データ表の寸法が大きくなると, またデータ構造が曖昧であるほど, 高い“寄与率”(大きい固有値)は期待できない. つまり少数次元内に収まらない.
- たとえば, 自由回答分析ではデータ表の寸法は数百～数千(ときに万単位)となる.
- 同時布置図内の点の重要度に濃淡があって, “見た目が同等”の情報量とはならない. 行(列)の要素の成分への“関与(寄与)”の程度が違うということ.
- よって, “寄与度”で複数成分を比べる必要がある.

計算対象となる「データ表」はなにか？

- 分析対象とする“データ表”の要素の記述に何通りかある.
- 見方を変えただけで、得られる結果は同等であることが分かっている. ソフトウェアに任せればよい.
- 厳密に言えば微妙な違いはある. たとえば得られる固有値, 固有ベクトルの出方など. 実用上は無視してよい.
- ここでの議論には必要ない情報なので略す. テキストに一部記述がある.
- 重要なことは, どのようなやり方でもストレッチ・プロファイル空間内の点(行あるいは列の要素)の分布のチラバリの程度を“総変動”(全慣性)で測ること.
- 行の側から(m 個の点)観察するとして, 以下のデータ行列を考えること. 列の側(n 個の点)は添字を入れ換え.

(つづき)

- 書籍, 文献, 研究者によって記述方法に違いはあるが, 言っていることは同じ.
- ①がストレッチ・プロファイルの空間で考えること.
- ②は①を平均のまわりで中心化したこと. ③は①の対称化. ④はカイ二乗統計量の要素を用いて記述.

$$\textcircled{1} \mathbf{X}_{m \times n} = (x_{ij}) = \left(\frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} \right) (i \in I, j \in J) \Rightarrow \mathbf{X}^* = (x_{ij}^*) \text{に同じ}$$

$$\textcircled{2} \mathbf{X}^* = (x_{ij}^*) = (x_{ij} - \bar{x}_j) = \left(\frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} - \sqrt{p_{+j}} \right) \Leftrightarrow \mathbf{V}_{n \times n} = (s_{jj'}) = (\mathbf{X}^*)^t \mathbf{P}_I \mathbf{X}^*$$

$$\textcircled{3} \mathbf{Q} = (y_{ij}) = \left(\frac{p_{ij}}{\sqrt{p_{i+} p_{+j}}} \right) = \left(\frac{f_{ij}}{\sqrt{f_{i+} f_{+j}}} \right) (i \in I, j \in J) \Leftrightarrow \mathbf{V}_{n \times n}^* = \mathbf{Q}^t \mathbf{Q} = \mathbf{P}_J^{-1/2} \mathbf{P}_{II} \mathbf{P}_I^{-1} \mathbf{P}_{IJ} \mathbf{P}_J^{-1/2}$$

$$\textcircled{4} \mathbf{Y}_{m \times n}^* = (y_{ij}^*) = \left(\frac{p_{ij} - p_{i+} p_{+j}}{\sqrt{p_{i+} p_{+j}}} \right) (i \in I, j \in J) \Leftrightarrow \mathbf{Y}_{m \times n}^* = (y_{ij}^*) = \mathbf{P}_I^{-1/2} (\mathbf{P}_{IJ} - \mathbf{r} \mathbf{c}^t) \mathbf{P}_J^{-1/2}$$

JMRAマーケティング・リサーチ講座

質的データのマイニング のための対応分析法

スライド資料[その3]

大隅 昇

ohsumi@ss.ij4u.or.jp

<http://wordminer.org/>

Copyright by Noboru Ohsumi

資料[その3]で述べること

- 同じ例を用いて若干の復習.
- 解釈に必要なとなる基本的な指標
- 数値例, 適用事例による観察
- 対応分析法の基本原理(仕組み)
- 仕組みの基本部分, 要点に絞って述べる(予定).
- 数式の細かい説明は控えて“**どう考えるか**”, “**なにを行っているのか**”を述べる.
- (フランス流の)“方言”の言い換えや補足説明.

キーワード

- 2元クロス表, 2元データの対応分析
- 成分スコアとその意味, 解釈
- プロファイルとストレッチ・プロファイル
- 三角座標系・重心座標系
- プロファイルのカイ二乗距離
- 成分スコアと合成変数(合成指標)
- 成分スコアの布置図, 同時布置図
- 成分スコア的双対性
- 総変動(全慣性)と固有値総和, 寄与率ほか
- 寄与度(絶対寄与度, 相対寄与度)

まず簡単な例で“仕組み”を観察

- すでに述べた事項を同じ例で再確認する.
 - クロス表の見方, 記法ほか
 - カイ二乗統計量との関係
- 「環境意識調査」の例を再び用いて, 対応分析法が行うことをざっとみる(すこし言い換える).
 - “数量化”とは何を行うのか
 - 質的データに与えられた“成分スコア”の意味
 - いくつかの“記号, 記法”に慣れる
- まず, “こんなことが目標らしい”と聞いていただく.

例による確認：環境意識調査から

- 再び「都市環境のすみやすさに関する調査」から引用.
- 2つの質問のクロス表(F表)を用いる.

$$\mathbf{F}_{4 \times 5} = (f_{ij})(i \in I, j \in J) \Leftarrow \mathbf{F} = (f_{ij})(i \in I, j \in J)_{m \times n}$$

質問I: あなたは、いま住んでいるまちが気に入っていますか.

1. たいへん気に入っている
2. まあ気に入っている
3. あまり気に入っていない
4. 気に入っていない

$$I = \{1, 2, 3, 4\} \Leftarrow I = \{1, 2, \dots, i, \dots, m\}$$

質問J: あなたが住んでいる地区は、都市としては、“緑(みどり)が多い”と感じますか. それとも少ないと感じますか.

1. かなり多い
2. 多いほうである
3. ふつう
4. 少ない
5. 少ないほうである

$$J = \{1, 2, 3, 4, 5\} \Leftarrow J = \{1, 2, \dots, j, \dots, n\}$$

ここでは、名義尺度・順序尺度;これだけではない一般化可能

はじめに, この例で一通り確認する

- いろいろな用語や式, 概念が登場する. ここでは, “**こんな用語や言葉**があるらしい”ことを知る.
- 用語・語句の厳密な定義はさておき, まずは“どういう仕組み”になっているか, を知ること.
- 仕組みを確認するために数式を若干用いる. “**数値例の確認**”で済ませ, 式が“**何を意味するか**”を説明する.
- “**トイ・データ**”あるいは簡単な“**実例データ**”で確認する(これが理解を容易にするはず).
- 同時に, ここで登場する用語・語句は基本として知っておくことが必要. “誤用”を避けるためには望ましいこと.

2元クロス表(F表): $\mathbf{F} = (f_{ij}) (i \in I, j \in J)$

このクロス表に“対応分析法”を適用する(対応分析法で解く).
このデータ表は“**多次元データ**”であること(行・列の両方向).

(I×J)クロス表		J = {1,2,3,4,5} 質問 J : 都市としては“緑(みどり)が多い”と感じますか. (n = 5)					
I = {1,2,3,4} 質問 I : いま住んでいるまちが気に入っていますか. (m = 4)	選択肢(j) → (i) ↓	1.かなり多い	2.多いほう	3.ふつう	4.少ないほう	5.少ない	行和 (f_{i+})
	1.大変気に入っている	166	239	86	26	7	524 f_{1+}
	2.まあ気に入っている	131	598 f_{ij}	324	146	36	1,235 f_{2+}
	3.あまり気に入っていない	6	40	55	51	20	172 f_{3+}
	4.気に入っていない	2	2	0	5	6	15 f_{4+}
	列和 (f_{+j})	305 f_{+1}	879 f_{+2}	465 f_{+3}	228 f_{+4}	69 f_{+5}	1,946 (N)

ここでは, $\sum_{i=1}^4 \sum_{j=1}^5 f_{ij} = \sum_{i=1}^4 f_{i+} = \sum_{j=1}^5 f_{+j} = N (=1,946)$

クロス表の見方[観察の向き]

- 2つの項目(質問文)のどちらから眺めるか.
- いま, 行(質問項目 I)の側から, この4つの選択肢が, 列(質問項目 J)の5つの選択肢の“5次元の空間”に布置すると考える(後述). 実際は4次元(であり3次元).
- つまり, クロス表とは“**多次元データ**”(多変量構造のデータ表と類似)であること, と考える.
- 「行」と「列」を, 入れ替えて読み替えても, **状況は同じ**である, と考える(“**対称性**”^(†)がある).
- この多次元データ表を, ある加工でえたデータ表の“**固有値問題**”あるいは“**特異値分解**”を行う.
- ここは, 形式的にこういうものだ, と考えておく(説明はあとで).

(†) “対称性”があるとするとは, クロス表の行と列の間の“**因果性**”を考えるモデル(向きがある)ではない, ということ.

この例で調べることを列記

◎1回目で述べたことに情報を追加して述べる.

- ①特異値・固有値, 寄与率と累積寄与率
- ②行・列の選択肢(標識)の“成分スコア”を求める
- ③第1成分スコアを布置図とし観察
- ④成分スコアの意味(回答選択肢との関係)
- ⑤“第1成分スコアの相関”を調べる
- ⑥第1成分スコアの並べ替え(クロス表との関係)
- ⑦統計量の性質を確認(要点のみ)

(つづき)

⑧重要な性質の確認

性質1: 平均値, 分散と固有値・特異値の意味

性質2: 固有値の総和とカイ二乗統計量の関係

⑨行と列の成分スコアの散布図(布置図)の観察

⑩「同時布置図」を描く(行と列の対応を観察)

- 繰り返すが, 登場する記号, 記法, 符丁をイメージとして確認する.
- “何を行っているのか”をおおまかに把握する.

①特異値・固有値，寄与率と累積寄与率

- 形式的には，“ある行列”の固有値問題・特異値分解を行うことで，必要な情報が得られる。
- 固有値，特異値，寄与率，累積寄与率を要約する。
- 固有値・特異値は大きさの順に並ぶ，
- ここで“特異値・固有値”とはなにか（何を意味するか）。
- 固有値総和はカイ二乗統計量と関係すること。
- なぜ，特異値・固有値はここでは3個なのか。
- 細かいことは後述する（繰り返し述べる）。

“ある行列”とは何か，その作り方がいくつかあること
しかし，それぞれが同等のこと（が多い）。

基本情報の要約

テキスト I 部, 15ページ

k	特異値 $\alpha_k = \sqrt{\lambda_k}$	固有値 λ_k	寄与率 ν_k	累積寄与率 $\sum_k \nu_k$	累積寄与率 (%)
1	α_1 0.35288	λ_1 0.12452	0.7121	0.7121	71.2
2	α_2 0.20959	λ_2 0.04393	0.2512	0.9633	96.3
3	α_3 0.08014	λ_3 0.00642	0.0367	1.0000	(100)
	固有値の総和 (全慣性)	0.17487 $\left(\sum_{k=1}^3 \lambda_k \right)$	1.000	—	—

- 固有値, 特異値, 寄与率, 累積寄与率
- この表の各数値と内容を次ページのように確認, 読む.

“形式的に”以下を確認する

- 固有値(分散, 慣性)の総和＝総変動(全慣性)となる.
- 総変動を, このクロス表の示す“情報の総量”と考える.
- クロス表の総度数(N)は, 表により異なるので, ピアソンのカイ二乗統計量をこれで除して“大きさを調整した量”とし, 総変動とする.
- ここでいう“情報”とは, 行あるいは列の個々の選択肢(標識)に付与の“成分スコアの分布の変動・チラバリの大きさ”をいう.

(つづき)

- (おおまかには)“配置図の点(成分スコア)の分布と変動”のこと.
- すでにいくつか例をみたことを思い出そう.
- 寄与率 = $[\text{固有値} / \text{固有値総和}] \times 100(\%)$ とする.
- 全体の情報量つまり総変動からみた, 個々の成分スコアの分散(固有値)の占める大きさを知る指標.
- 累積寄与率は, 寄与率の累積値となる.

数値で確認する

①ピアソンのカイ二乗統計量: $\chi_p^2 = 340.309$

②固有値の総和 = $\frac{\text{ピアソンのカイ二乗統計量}}{\text{2元データ表の総度数}}$ の関係を確認

$$\sum_{k=1}^K \lambda_k \Rightarrow \sum_{k=1}^3 \lambda_k = 0.17487 \Leftrightarrow \frac{\chi_p^2}{N} = \frac{340.309}{1946} = 0.174876 \dots$$

これを, このクロス表の情報の総量つまり“総変動(全慣性)”と読む.

③ここで, 固有値の数は, $K = \min\{m, n\} - 1$ となる.

つまり, $K = \min\{4, 5\} - 1 = 3$

固有値数は行数, 列数の小さい値から1を引いた数だけある.

また, 固有値は非負で1を越えない. つまり, $0 \leq \lambda \leq 1$ となる.

$$\phi^2 = \frac{\chi_p^2}{N}: \text{平均平方関連係数という}; \phi = \sqrt{\frac{\chi_p^2}{N}}: \text{ファイ係数という}$$

(つづき)

④固有値(分散)の大きさは, 添字の大きさの順に対応とする.

つまり, $\lambda_1 > \lambda_2 > \lambda_3$ の順となる(表で確認).

成分スコアの分散は次第に小さくなる.

⑤個々の固有値は各成分スコアの“分散(慣性)”である.

⑥特異値は固有値の正の平方根である(表で確認).

- ここでは, $\alpha_k = \sqrt{\lambda_k}$ と記す. ここで, $0 \leq \alpha_k \leq 1$ となる.
- これは, 成分スコアの“標準偏差”である.
- 行と列の対応する成分スコアの“相関係数”でもある.

②行・列の選択肢の“成分スコア”の確認

		$\lambda_1 = 0.12452 \quad \lambda_2 = 0.04393 \quad \lambda_3 = 0.00642$ $\Downarrow \quad \quad \quad \Downarrow \quad \quad \quad \Downarrow$				
質問文	質問選択肢 (i)	第1成分スコア	第2成分スコア	第3成分スコア	z_{ik}	
		z_{i1}	z_{i2}	z_{i3}		
$I = \{1, 2, 3, 4\}$ 質問I あなたは、いま住んでいる まちが気に入って いますか。	1.大変気に入っている	-0.4442	0.2027	-0.0353	z_{ik}	行成分スコア
	2.まあ気に入っている	0.0623	-0.1315	0.0311		
	3.あまり気に入っていない	0.7886	0.1907	-0.1698		
	4.気に入っていない	1.3458	1.5567	0.6157		
$J = \{1, 2, 3, 4, 5\}$ 質問J あなたの住んでいる地区 は、都市としては“緑 (みどり)が多い”と感 じますか。	質問選択肢 (j)	z_{j1}^*	z_{j2}^*	z_{j3}^*	z_{jk}^*	列成分スコア
	1. かなり多い	-0.5403	0.3235	-0.0640		
	2.多いほう	-0.1118	-0.1055	0.0657		
	3.ふつう	0.1545	-0.1506	-0.0613		
	4.少ないほう	0.5530	0.0750	-0.1069		
	5.少ない	0.9438	0.6805	0.2119		

成分スコア(数量化得点)とはどのようにして得られたのか

もとのクロス表(確率行列)と成分スコアの関係

$$z_{ik} \ (i \in I), z_{jk}^* \ (j \in J)$$

$$\left(\begin{array}{l} k = 1, 2, \dots, K \\ K = \min\{m, n\} - 1 \end{array} \right)$$

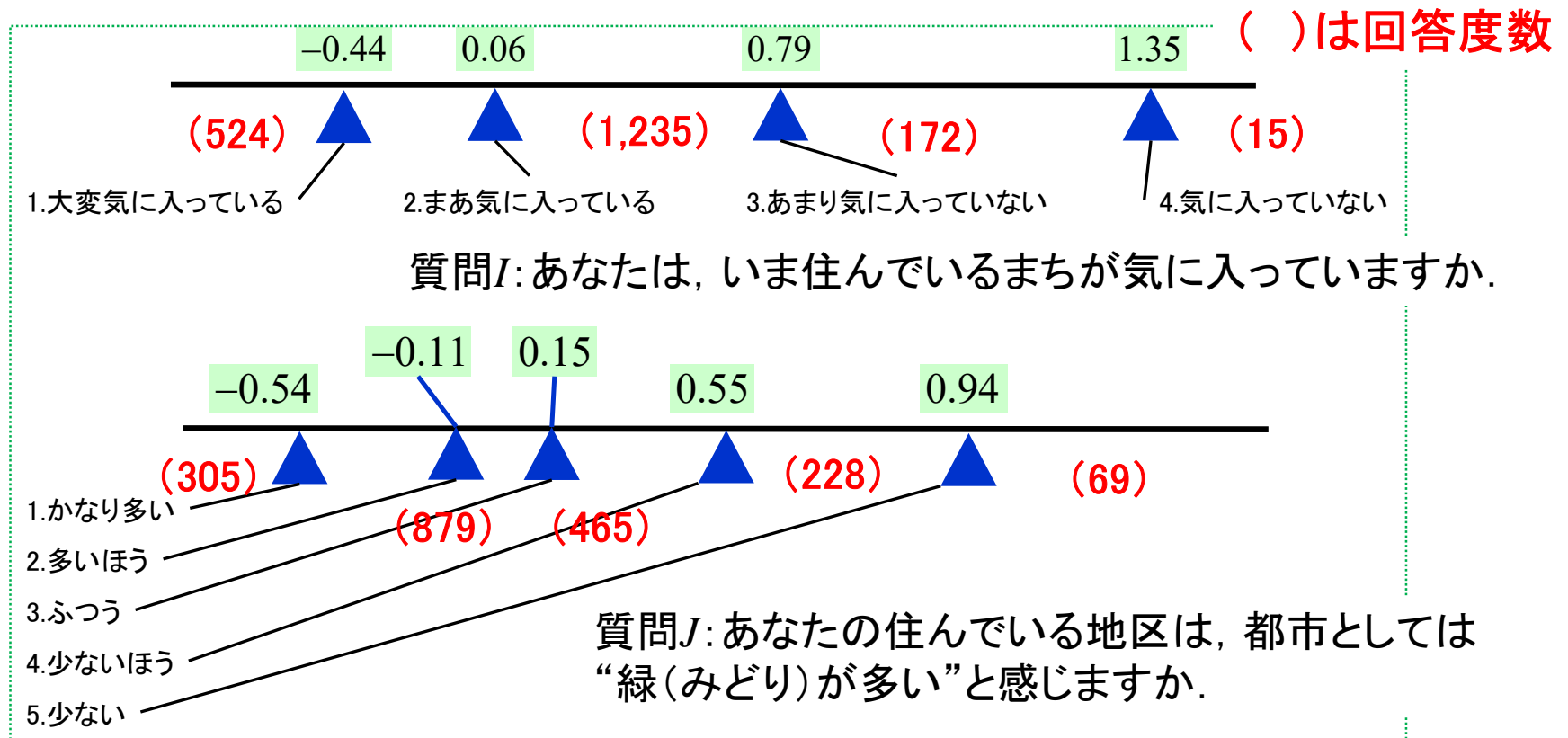
表 16 項目 I, J の選択肢の成分スコアと確率行列の関係

		項 目 J						成分スコア \mathbf{Z} $m \times K$							
		1	2	...	j	...	n	1	2	...	k	...	k'	...	K
項 目 I	1	p_{11}	p_{12}	...	p_{1j}	...	p_{1n}	z_{11}	z_{12}	...	z_{1k}	...	$z_{1k'}$...	z_{1K}
	2	p_{21}	p_{22}	...	p_{2j}	...	p_{2n}	z_{21}	z_{22}	...	z_{2k}	...	$z_{2k'}$...	z_{2K}
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	i	p_{i1}	p_{i2}	...	p_{ij}	...	p_{in}	z_{i1}	z_{i2}	...	z_{ik}	...	$z_{ik'}$...	z_{iK}
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	m	p_{m1}	p_{m2}	...	p_{mj}	...	p_{mn}	z_{m1}	z_{m2}	...	z_{mk}	...	$z_{mk'}$...	z_{mK}
成 分 ス コ ア	1	z_{11}^*	z_{21}^*	...	z_{j1}^*	...	z_{n1}^*	<div>行の項目 I の選択肢の成分スコア</div> $\mathbf{Z} = \mathbf{P}_I^{-1} \underbrace{\mathbf{P}_{IJ} \mathbf{P}_J^{-1/2}}_{m \times n} \mathbf{L}$ $\begin{matrix} m \times K & & m \times n & & n \times K \end{matrix}$ <div>列の項目 J の選択肢の成分スコア</div> $\mathbf{Z}^* = \mathbf{P}_J^{-1} \underbrace{\mathbf{P}_{JI} \mathbf{P}_I^{-1/2}}_{n \times m} \mathbf{U}$ $\begin{matrix} n \times K & & n \times m & & m \times K \end{matrix}$ <div>ここで $K = \min\{m, n\} - 1$</div>							
	2	z_{12}^*	z_{22}^*	...	z_{j2}^*	...	z_{n2}^*								
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots								
	k	z_{1k}^*	z_{2k}^*	...	z_{jk}^*	...	z_{nk}^*								
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots								
	k'	$z_{1k'}^*$	$z_{2k'}^*$...	$z_{jk'}^*$...	$z_{nk'}^*$								

この対応を覚えよう!
(項目と選択肢, 記法)

③第1成分スコアを布置図とする

- “第1成分スコア”に注目, これについて“布置図”を描く.



確認

- 固有値の大きい(成分スコアの分散が大きい＝情報が多い)ほうから, 調べること. 布置図を作ること.
- もとの選択肢に付与の“コード”(数値)または“標識”とは異なる“数量(成分スコア)”を得たこと.
- つまり“質的データ”(選択肢という標識)にあらたな数量を付与した(これは加減乗除が可能). 数量化できた.
- この成分スコアは, この調査の回答者の意識態度のある側面を(総合的な“尺度”として)数量化できたと考える.

成分スコア(coordinates)は, 主座標, 座標, スコア, 数量化得点とさまざまな呼称がある.

(つづき)

- もとの選択肢に(形式的に)付与の数値コードの並びと, “たまたまこの例では合っている”ということ.
- ここでは質問文選択肢の作り方がうまく機能したこと.
- 目標の1つに, “なるべく(序列)尺度として機能”するような質問項目は作れるか, がある. 再現性, 反復利用.
例: 住みやすさ感, 生活満足感など
- もとのクロス表あるいはさらに遡ってもとのデータ表と比べる.

もとのデータ表(多変量構造)

- 各回答者の、もとの選択肢(という標識, 質的情報)に, 成分スコアという数量が付与された. [第 I 部, 8ページあたり]

★★SURVEY83.DAT.edit(1973cases)_new - JMP

ファイル(F) 編集(E) テーブル(T) 行(R) 列(C) 実験計画 (DOE)(D) 分析(A) グラフ(G) ツール(O) アドイン(N) 表示(V) ウィンドウ(W) ヘルプ(H)

列(127/2)	す	近くの緑地や公園に、 どのくらい出かけま...	近くの緑地や公園に、どのくら い出かけますか。	いま住んでいるまちが気に入っていますか。	住んでいる地区は緑(みどり)が多いと感じ ますか。
1	56	1	1.毎日のように行く	1.大変気に入っている	2.多いほう
2	53	3	3.月に1~2回くらい	2.まあ気に入っている	2.多いほう
3	42	3	3.月に1~2回くらい	2.まあ気に入っている	2.多いほう
4	56	3	3.月に1~2回くらい	1.大変気に入っている	1.かなり多い
5	53	1	1.毎日のように行く	2.まあ気に入っている	2.多いほう
6	42	2	2.週に1~2回くらい	2.まあ気に入っている	2.多いほう
7	54	1	1.毎日のように行く	2.まあ気に入っている	2.多いほう
8	42	3	3.月に1~2回くらい	2.まあ気に入っている	2.多いほう
9	47	4	4.年に1~2回くらい	2.まあ気に入っている	2.多いほう
10	54	3	3.月に1~2回くらい	2.まあ気に入っている	1.かなり多い
11	56	2	2.週に1~2回くらい	1.大変気に入っている	1.かなり多い
12	42	5	5.ほとんど出かけない	2.まあ気に入っている	3.ふつう
13	50	4	4.年に1~2回くらい	2.まあ気に入っている	1.かなり多い
14	54	2	2.週に1~2回くらい	1.大変気に入っている	1.かなり多い
15	54	3	3.月に1~2回くらい	2.まあ気に入っている	2.多いほう
16	42	2	2.週に1~2回くらい	1.大変気に入っている	1.かなり多い
17	57	3	3.月に1~2回くらい	2.まあ気に入っている	1.かなり多い
18	44	5	5.ほとんど出かけない	1.大変気に入っている	2.多いほう
19	42	3	3.月に1~2回くらい	2.まあ気に入っている	1.かなり多い
20	46	2	2.週に1~2回くらい	2.まあ気に入っている	2.多いほう
21	54	4	4.年に1~2回くらい	2.まあ気に入っている	2.多いほう
22	99	3	3.月に1~2回くらい	1.大変気に入っている	2.多いほう
23	54	2	2.週に1~2回くらい	3.あまり気に入っていない	3.ふつう

すべての行 1,973
選択されている行 24
除外されている行 27
表示しない行 0
ラベルのついた行 0

④成分スコアの意味(回答選択肢との関係)

回答 パターン	質問I	質問J	回答数 (人)	行の第1成分スコア z_{i1}	列の第1成分スコア z_{j1}^*
1	1.大変気に入っている	1.かなり多い	166	-0.4442	-0.5403
2	1.大変気に入っている	2.多いほう	239	-0.4442	-0.1118
3	1.大変気に入っている	3.ふつう	86	-0.4442	0.1545
4	1.大変気に入っている	4.少ないほう	26	-0.4442	0.5530
5	1.大変気に入っている	5.少ない	7	-0.4442	0.9438
6	2.まあ気に入っている	1.かなり多い	131	0.0623	-0.5403
7	2.まあ気に入っている	2.多いほう	598	0.0623	-0.1118
8	2.まあ気に入っている	3.ふつう	324	0.0623	0.1545
			146	0.0623	0.5530
			36	0.0623	0.9438
			6	0.7886	-0.5403
			40	0.7886	-0.1118
13	3.あまり気に入っていない	3.ふつう	55	0.7886	0.1545
14	3.あまり気に入っていない	4.少ないほう	51	0.7886	0.5530
15	3.あまり気に入っていない	5.少ない	20	0.7886	0.9438
16	4.気に入っていない	1.かなり多い	2	1.3458	-0.5403
17	4.気に入っていない	2.多いほう	2	1.3458	-0.1118
18	4.気に入っていない	3.ふつう	0	—	—
19	4.気に入っていない	4.少ないほう	5	1.3458	0.5530
20	4.気に入っていない	5.少ない	6	1.3458	0.9438

成分スコアは「量的データ」である

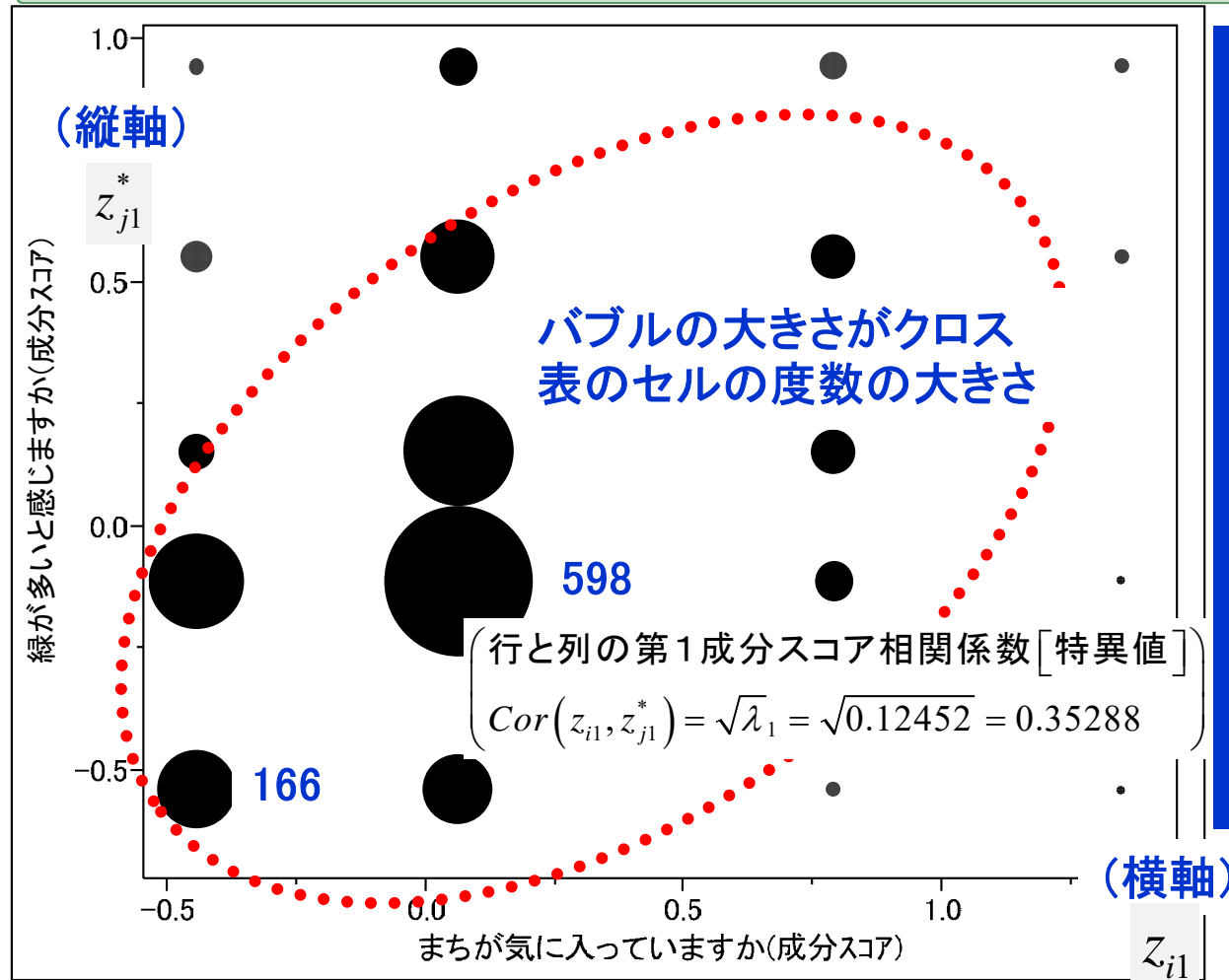
- 「第1成分スコア」のとき.
- もとのクロス表の度数の書き替え
- クロス表のセル数は“20通り”

データ表で確認

©Noboru Ohsumi

⑤「第1成分スコア」の相関を調べる

まちが気に入っていますか(成分スコア)と緑が多いと感じますか(成分スコア)のバブルプロット サイズ:回答数



0.94 (305)

0.55 (879)

0.15 (465)

-0.11 (228)

-0.54 (69)

()内回答度数

円のサ

-0.44

0.06

0.79

1.35

(524)

(1,235)

(172)

(15)

次のような関係にあると読む

第 1 成分の固有値 [第 1 成分スコアの分散 (慣性)]

$$\lambda_1 = 0.12452$$

第 1 成分スコアの特異値 [第 1 固有値の正の平方根]

$$\alpha_1 = \sqrt{\lambda_1} = \sqrt{0.12452} = 0.35288$$

行と列の第 1 成分スコアの相関係数 [特異値]

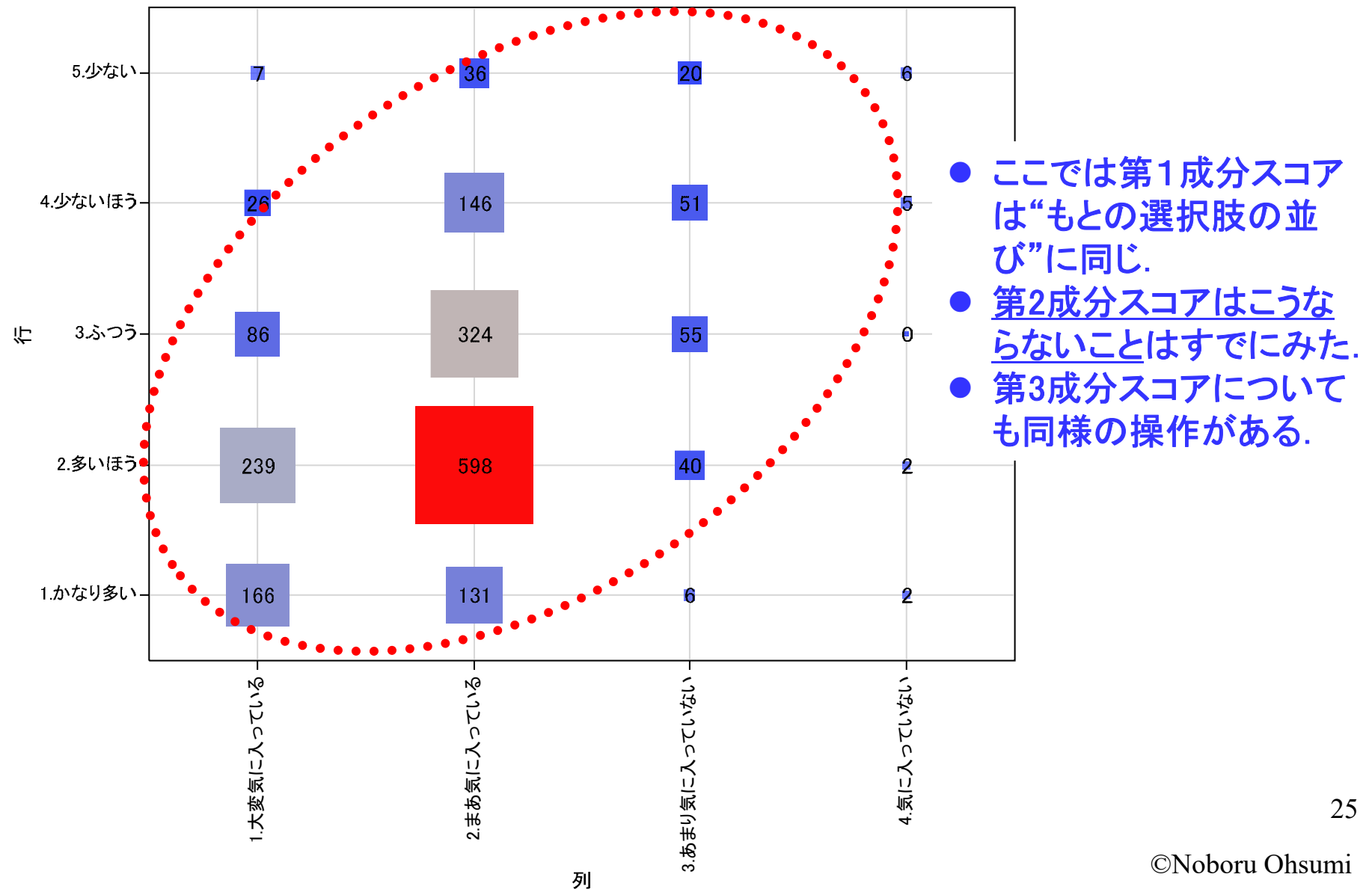
$$Cor(z_{i1}, z_{j1}^*) = 0.35288$$

ここで,

行の第 1 成分スコア : z_{i1} ($i = 1, 2, \dots, 4$),

列の第 1 成分スコア : z_{j1}^* ($j = 1, 2, \dots, 5$)

⑥第1成分スコアの並べ替え(クロス表対応)



⑦統計量の性質を確認(要点のみ)

- 成分スコアの平均値, 分散についての性質.
- 平均値, 分散(固有値), 標準偏差の算出と意味.
- 成分スコア間の相関の意味.
- 2元データ表の情報の総量(総変動)を知る.
- “ピアソンのカイ二乗統計量”と“総変動(全慣性)”
- “固有値”とその総和(総変動)の関係.

分散(慣性, inertia)の和と総変動・全慣性(total inertia),
特異値, 相関係数

これらを式で示す(誘導, 例で確認)

行の選択肢 $i \in I$ の第 k 成分スコアの平均値

$$\bar{z}_k = \frac{1}{N} \sum_{i=1}^m f_{i+} z_{ik} = \sum_{i=1}^m p_{i+} z_{ik} = 0 \quad (k = 1, 2, \dots, K)$$

列の選択肢 $j \in J$ の第 k 成分スコアの平均値

$$\bar{z}_k^* = \frac{1}{N} \sum_{j=1}^n f_{+j} z_{jk}^* = \sum_{j=1}^n p_{+j} z_{jk}^* = 0 \quad (k = 1, 2, \dots, K)$$

「0」である
そうなるよう
に調整した

行の選択肢 $i \in I$ の第 k 成分スコアの分散と標準偏差

$$V[z_{ik}] = \lambda_k; \sqrt{V[z_{ik}]} = \sqrt{\lambda_k} = \alpha_k \quad (k = 1, 2, \dots, K)$$

列の選択肢 $j \in J$ の第 k 成分スコアの分散と標準偏差

$$V[z_{jk}^*] = \lambda_k; \sqrt{V[z_{jk}^*]} = \sqrt{\lambda_k} = \alpha_k \quad (k = 1, 2, \dots, K)$$

どちらも等しい
(同じであること
に注意)

⑧性質1: 平均値, 分散と固有値・特異値

- 行成分スコア, 列成分スコアのいずれについても, 以下の性質がある.
 - i) 成分スコアの“**平均値は「ゼロ」**”である.
あるいはそうなるように調整した[そうではない場合もあるが基本はこれ]. 平均値で中心化した.
 - ii) 成分スコアの“**分散(慣性)は固有値**”に等しい.
 - iii) よって, 成分スコアの“**標準偏差は特異値**”に等しい.
 - iv) 特異値は行成分スコアと列成分スコアの“**相関係数**”.
- 成分スコア(数量化得点)の算出についてはあとで述べる. ここでまず“数値例”で上の性質を調べる.

成分スコアの“平均値”と“分散”の確認

$\lambda_1 = 0.124$ $\lambda_2 = 0.044$ $\lambda_3 = 0.006$ テキスト, 16ページ

質問 <i>I</i> と 選択肢	第1成分スコア	第2成分スコア	第3成分スコア	周辺和
	z_{i1}	z_{i2}	z_{i3}	f_{i+}
1.大変気に入っている	-0.4442	0.2027	-0.0353	524
2.まあ気に入っている	0.0623	-0.1315	0.0311	1,235
3.あまり気に入っていない	0.7886	0.1907	-0.1698	172
4.気に入っていない	1.3458	1.5567	0.6157	15
質問 <i>J</i> と 選択肢	z_{j1}^*	z_{j2}^*	z_{j3}^*	f_{+j}
1. かなり多い	-0.5403	0.3235	-0.0640	305
2. 多いほう	-0.1118	-0.1055	0.0657	879
3. ふつう	0.1545	-0.1506	-0.0613	465
4. 少ないほう	0.5530	0.0750	-0.1069	228
5. 少ない	0.9438	0.6805	0.2119	69

(行和)

(列和)



成分スコア(数量化得点)の算出については後述.

成分スコアの“平均値”の算出例

表 11 成分スコアの平均値の算出例

		z_{ik} の平均値 \bar{z}_k の算出		
	k i	第 1 成分	第 2 成分	第 3 成分
$f_{i+} z_{ik}$ の値	1	-232.7608	106.2148	-18.4972
	2	76.9405	-162.4025	38.4085
	3	135.6392	32.8004	-29.2056
	4	20.1870	23.3505	9.2355
平均値	\bar{z}_k	0.0000	0.0000	0.0000

		z_{jk}^* の平均値 \bar{z}_k^* の算出		
	k j	第 1 成分	第 2 成分	第 3 成分
$f_{+j} z_{jk}^*$ の値	1	-164.7915	98.6675	-19.5200
	2	-98.2722	-92.7345	57.7503
	3	71.8425	-70.0290	-28.5045
	4	126.0840	17.1000	-24.3732
	5	65.1222	46.9545	14.6211
平均値	\bar{z}_k^*	0.0000	0.0000	0.0000

$$\bar{z}_k = \frac{1}{N} \sum_{i=1}^m f_{i+} z_{ik} = 0$$

$$\bar{z}_k^* = \frac{1}{N} \sum_{j=1}^n f_{+j} z_{jk}^* = 0$$

テキスト(第 I 部), 16ページから引用
「成分スコア」の平均値=0である

成分スコアの“分散”の算出例

表 12 成分スコアの分散の算出例

		z_{ik} の分散算出		
	$j \backslash k$	第 1 成分	第 2 成分	第 3 成分
$\lambda_k = V[z_{ik}] = \frac{1}{N} \sum_{i=1}^m f_{i+} z_{ik}^2$ $f_{i+} z_{ik}$ の値	1	103.39235	21.52974	0.65295
	2	4.79339	21.35593	1.19450
	3	106.96507	6.25504	4.95911
	4	27.16766	36.34972	5.68630
分散	$V[z_{ik}]$	0.12452	0.04393	0.00642
		z_{jk}^* の分散算出		
	$j \backslash k$	第 1 成分	第 2 成分	第 3 成分
$\lambda_k = V[z_{jk}^*] = \frac{1}{N} \sum_{j=1}^n f_{+j} (z_{jk}^*)^2$ $f_{+j} (z_{jk}^*)^2$ の値	1	89.03685	31.91894	1.24928
	2	10.98683	9.78349	3.79419
	3	11.09967	10.54637	1.74733
	4	69.72445	1.28250	2.60550
	5	61.46233	31.95254	3.09821
分散	$V[z_{jk}^*]$	0.12452	0.04393	0.00642

固有値
= 分散 (慣性)

$$\lambda_1 = 0.12452$$

$$\lambda_2 = 0.04393$$

$$\lambda_3 = 0.00642$$

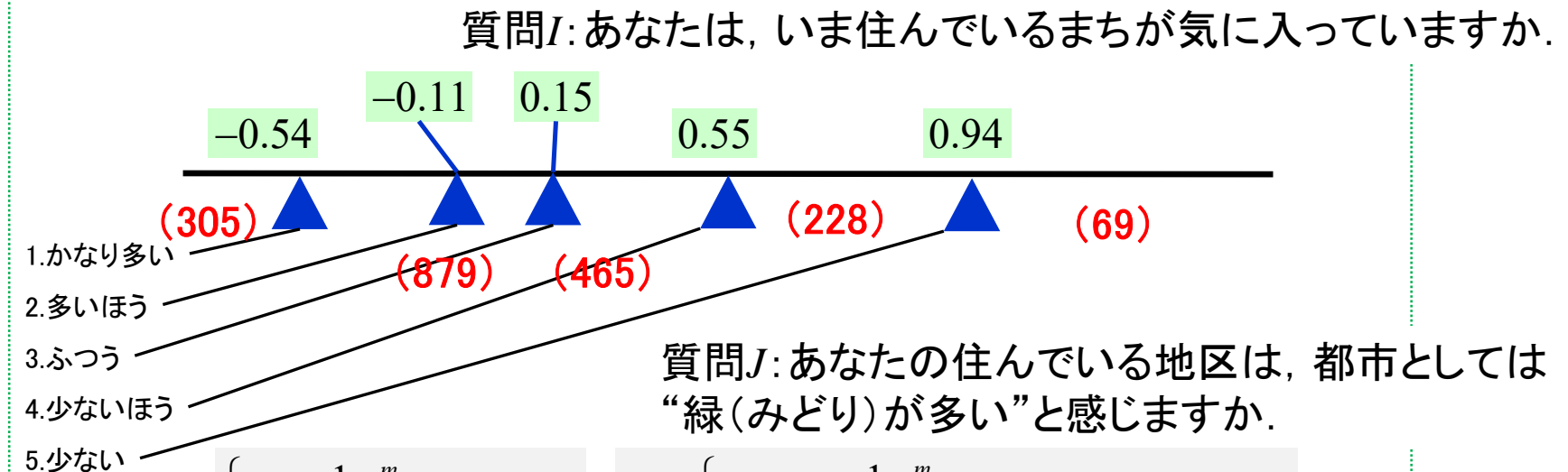
テキスト (第 I 部), 17 ページから引用
成分スコアの分散は固有値 (λ) である

$$V[z_{ik}] = \frac{1}{N} \sum_{i=1}^m f_{i+} \left(z_{ik} - \underbrace{\bar{z}_k}_{=0} \right)^2 = \frac{1}{N} \sum_{i=1}^m f_{i+} z_{ik}^2$$

$$V[z_{jk}^*] = \frac{1}{N} \sum_{j=1}^n f_{+j} \left(z_{jk}^* - \underbrace{\bar{z}_k^*}_{=0} \right)^2 = \frac{1}{N} \sum_{j=1}^n f_{+j} (z_{jk}^*)^2$$

第1成分スコアで再確認

- “第1成分スコア”の平均値と分散・慣性(固有値)を再チェック.



$$\begin{cases} \bar{z}_k = \frac{1}{N} \sum_{i=1}^m f_{i+} z_{ik} = 0 \\ \bar{z}_k^* = \frac{1}{N} \sum_{j=1}^n f_{+j} z_{jk}^* = 0 \end{cases}$$

$$\lambda_1 = \begin{cases} V[z_{i1}] = \frac{1}{N} \sum_{i=1}^m f_{i+} z_{i1}^2 = 0.12452 \\ V[z_{j1}^*] = \frac{1}{N} \sum_{j=1}^n f_{+j} (z_{j1}^*)^2 = 0.12452 \end{cases}$$

⑧性質2: 固有値の総和, カイ二乗統計量の関係

- i) 成分スコアの分散の総和, “総変動(全慣性)”は, 固有値の総和に等しい.
 - ii) カイ二乗統計量を総度数(N)で割った量と固有値の総和は等しい. クロス表の大きさ(つまり総変動, 全慣性)を調整.
- 以上を式で表すと以下のようになる.

成分スコアの分散の総和(総変動), 固有値の総和と全慣性の関係

$$\sum_{k=1}^K V[z_{ik}] = \sum_{k=1}^K V[z_{jk}^*] = \sum_{k=1}^K \lambda_k = \frac{\chi_p^2}{N} \quad (\text{ここで, } K = \min\{m, n\} - 1)$$

カイ二乗統計量と固有値の総和の関係

$$\frac{(\text{カイ二乗統計量})}{N} = \frac{\chi_p^2}{N} = \sum_{k=1}^K \lambda_k$$

数値例で再度確認すると, ...

固有値(分散)の総和(全慣性)の確認

ここで, $K = \min\{m, n\} - 1$ から

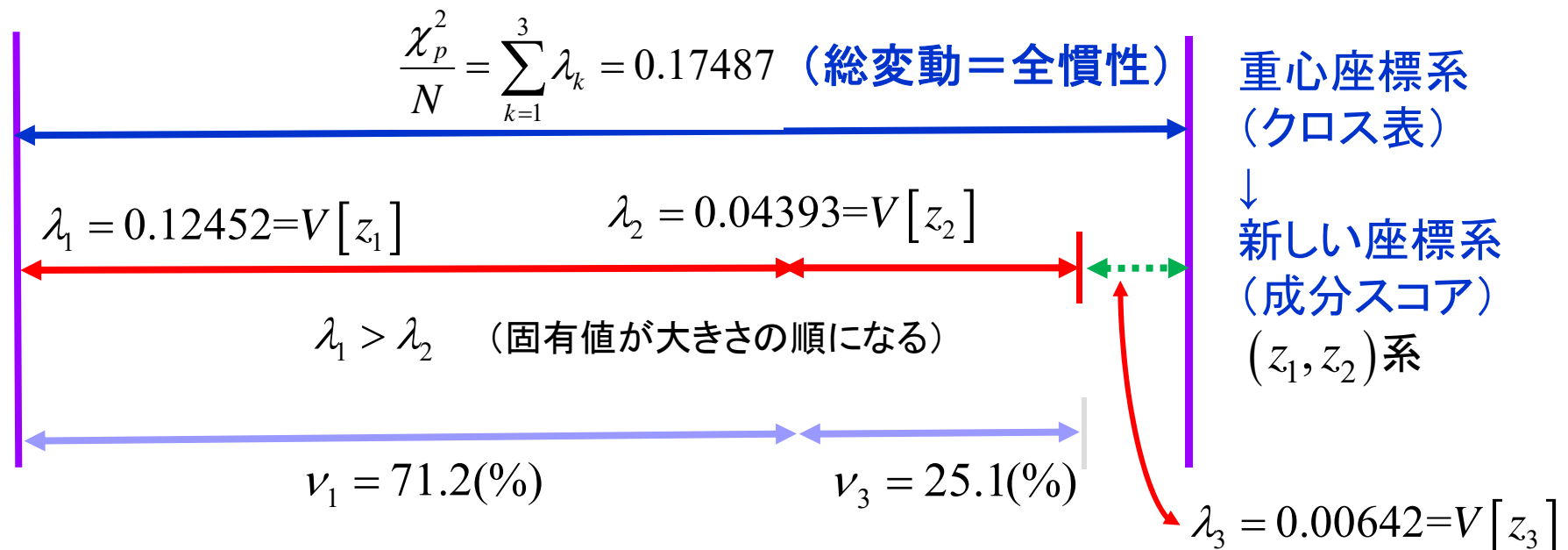
例では, $K = \min\{5, 4\} - 1 = 3$ となる.

$$\begin{aligned}\sum_{k=1}^3 V[z_{ik}] &= \sum_{k=1}^3 V[z_{jk}^*] = \sum_{k=1}^3 \lambda_k \\ &= \lambda_1 + \lambda_2 + \lambda_3 = 0.12452 + 0.04393 + 0.00642 = 0.17487\end{aligned}$$

カイ二乗統計量と固有値の関係

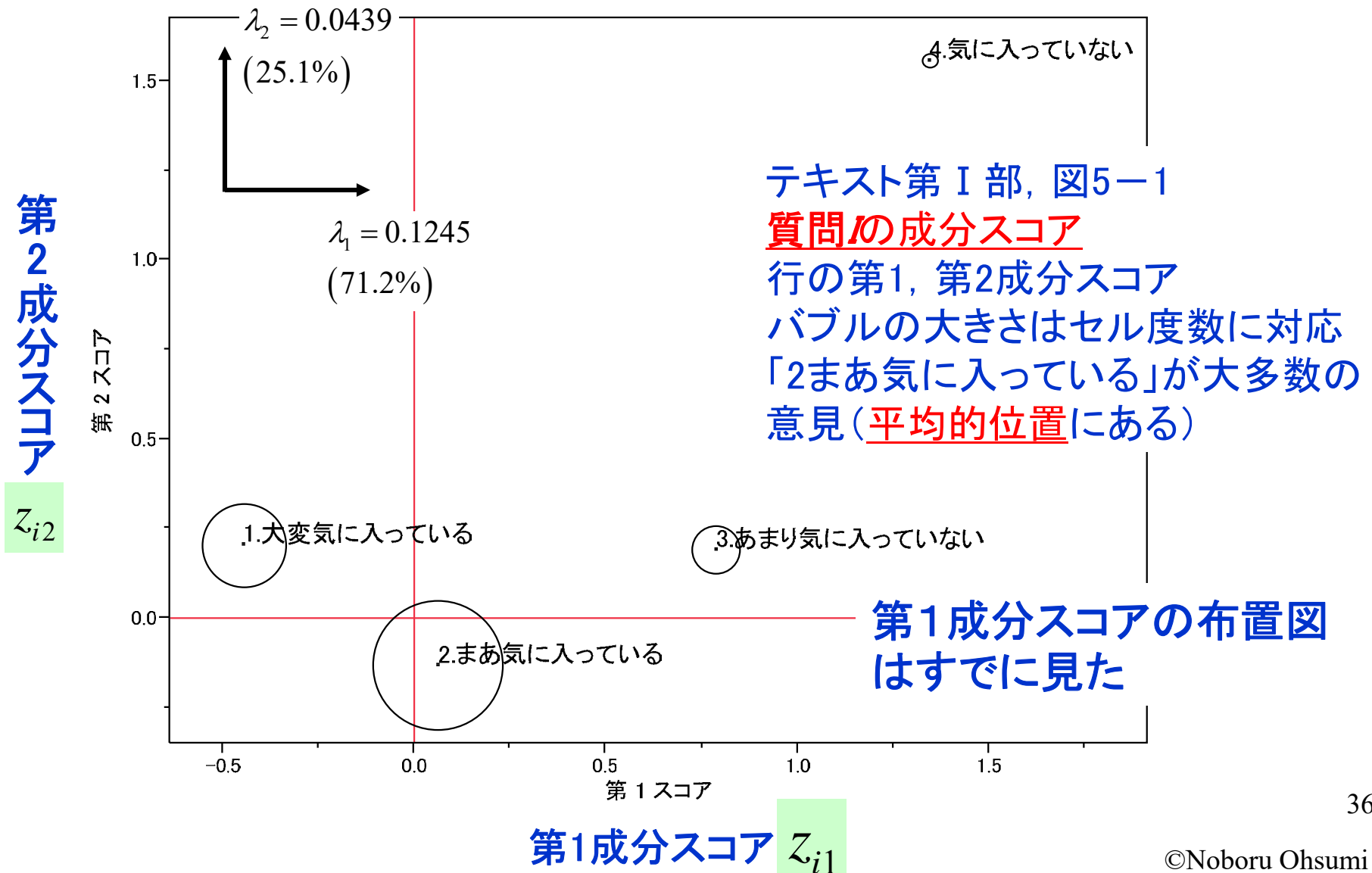
$$\begin{cases} \sum_{k=1}^3 \lambda_k = 0.174874 \Leftrightarrow \lambda_k = V[z_k] \text{と略記しておく} \\ \frac{\chi_p^2}{N} = \frac{340.309}{1946} = 0.174876\cdots = 0.17488 \end{cases}$$

総変動(慣性)と3つの固有値(分散)の関係

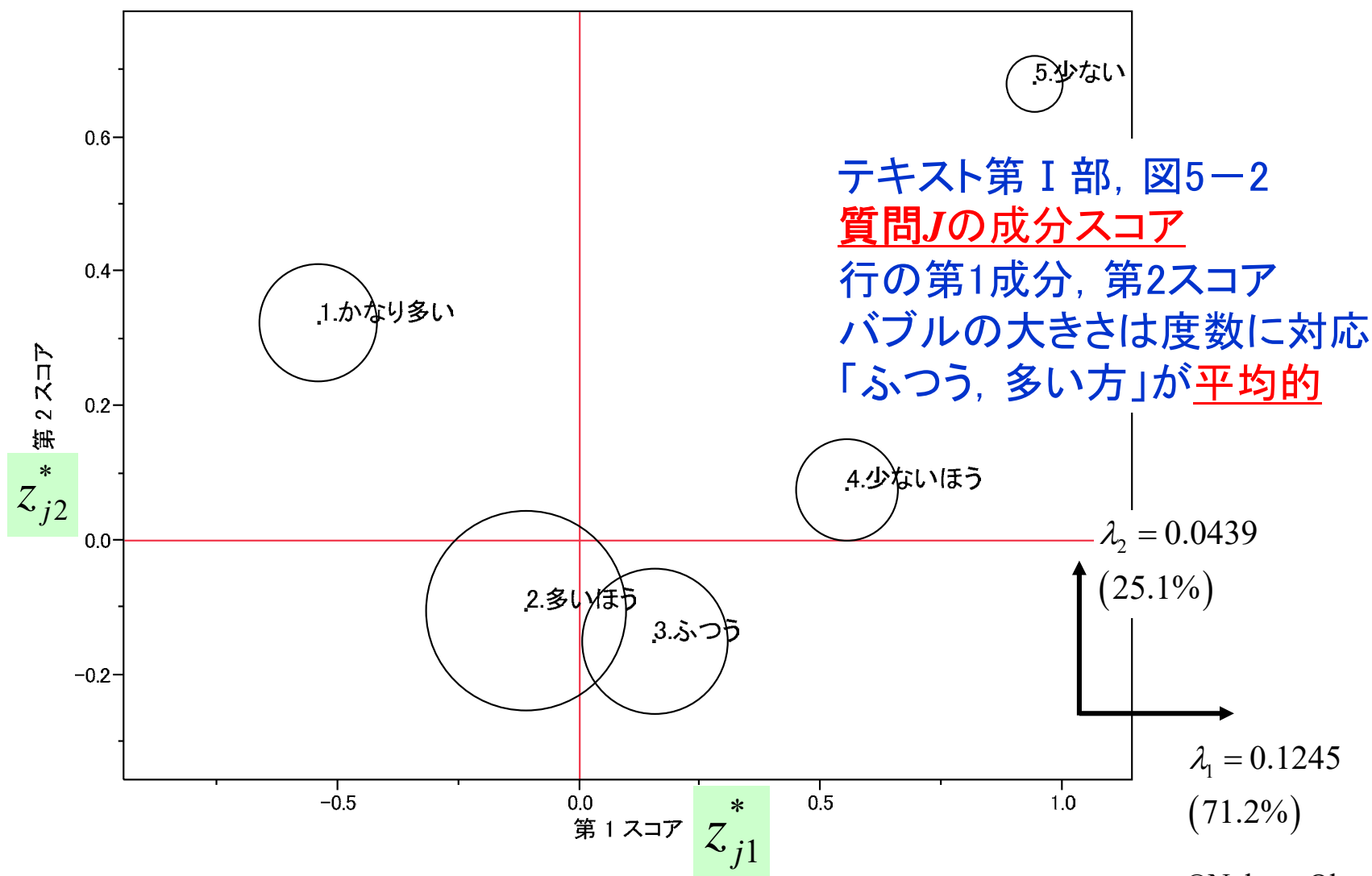


添字が“固有値の大きさ”に対応する(固有値=分散, 慣性)
 総変動を個々の成分スコアの分散がどのように分割するか
 ⇒寄与の大きさ(寄与率)として測る
形式的にははじめの1成分で約71%, 2成分で約96%の情報量となる.

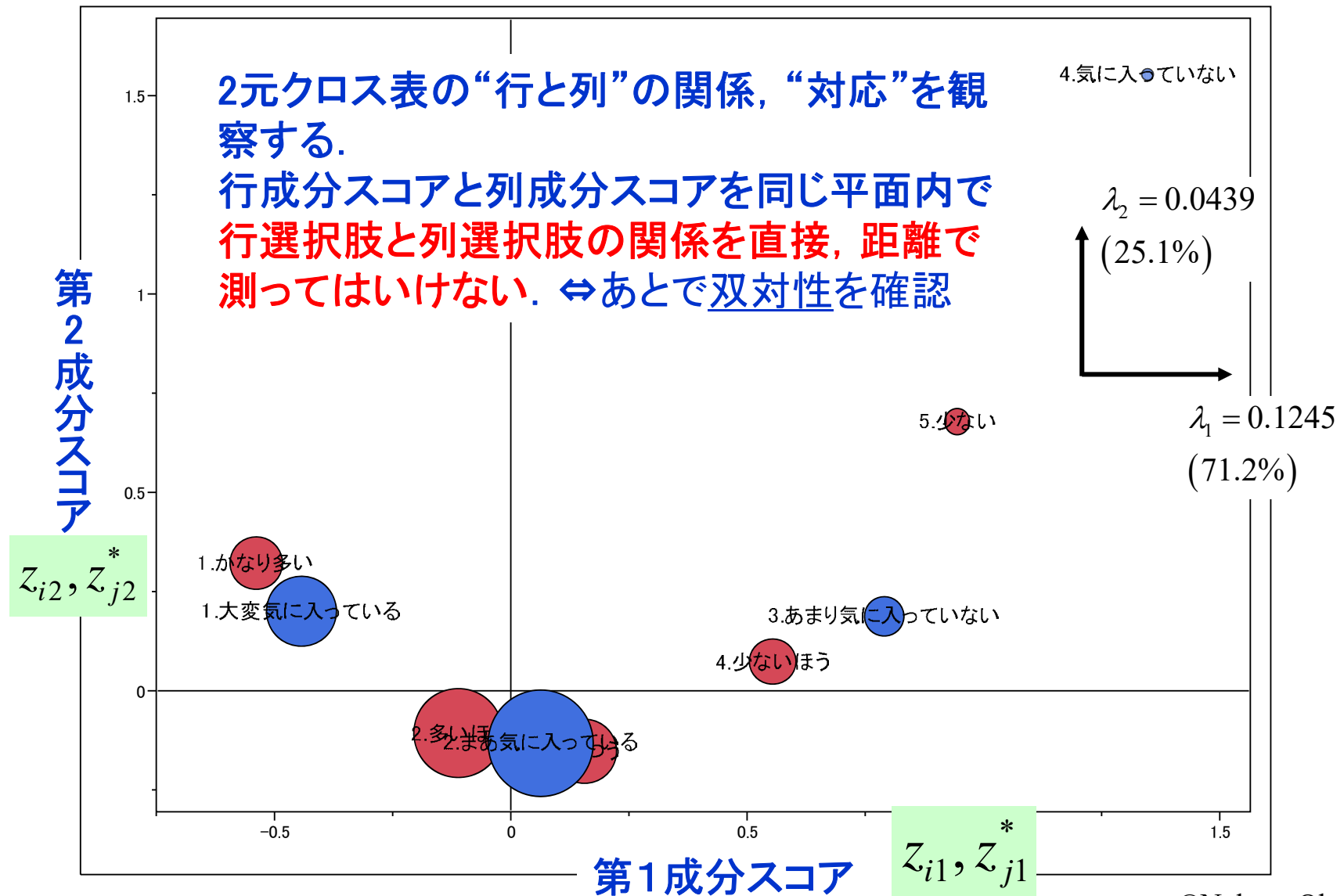
⑨行成分スコアの散布図(布置図)の観察



列成分スコアの散布図



⑩「同時布置図」を描く(行と列の対応を観察)



ここでなにを観察したか

- “数量化”とはなにか, が少しみえただろう.
- 項目(質問文)の行・列に与えられた“選択肢・標識は質的データ”である.
- はじめに付与の数値(コード)は, 名目的な情報で, 数値・数量としての量的データではない.
- “2元データ表”として, 選択肢を質的データという分類区分として見直す.
- その選択肢・標識に, 新たな“数量(成分スコア)”を付与した.
- これは四則演算が可能な量的データである.

対応分析では, 選択肢をカテゴリー, モダリティなどともいう.

(つづき)

- 尺度化(スケーリング)の発想:

量的データとみなせるような数量(スコア)が“かりに付与できた”として, 行選択肢と列選択肢のそのスコアの相関を最大化(最適化)すると考える. これが“林の数量化”(quantification)の考え方. 態度尺度の生成.

- 対応分析法:

2元データ表を多次元データとみて, 行・列のプロファイル(比率データ)を考え, この分布空間内で次元縮小化を図る. 結果として, 選択肢・標識にスコアを付与した. これがベンゼクリ流の考え方.

あらためて対応分析法を調べる

- まず、きわめて形式的に記述してみる.
- 対応分析法とは, ある行列の“固有値問題”となること.
- “特異値分解”でも説明できること.
- この意味で他の多変量解析手法との類似性がある.
- とくに, “主成分分析”とその類似手法, つまり“合成指標型・合成変数型”の手法との類似性.
- さらに(次元縮約を伴う)判別分析, 正準相関分析など.
- 一部の尺度化法(ガットマン尺度, 数量化法III類, その変形). 言葉として知っておこう.

(つづき)

- ここで“ある行列とは”はなにか？
- 成分スコアつまり“合成変数”(合成指標)を作ること.
- 合成変数・合成指標とはなにか？
- その合成変数としての“成分スコア”の形を示す. それをいきなり式で書いてみよう.
- ここは, こういうもの, と見ておく. 順次, 要約説明する.

数理的にはある行列の”特異値分解“を行うこと, あるいはその行列の積(正確には共分散行列)の”固有値問題“を解くこと.

特異値分解(SVD: singular value decomposition)

固有値問題(eigenvalue problem)と固有方程式

“合成変数(合成指標)”とはなにか

- 成分スコアの形(成分式)を具体的に眺めてみる.
- この式は, $q_{ij}/\sqrt{p_{+j}}, q_{ij}^*/\sqrt{p_{i+}}$ の“加重和”の形になっている.
- ではここで $q_{ij}/\sqrt{p_{+j}}$ あるいは $q_{ij}^*/\sqrt{p_{i+}}$ とはなにか?

項目 I の第 i 選択肢の第 k 成分スコア

$$z_{ik} = \sum_{j=1}^n l_{jk} \underline{x_{ij}} = \sum_{j=1}^n \left(\frac{p_{ij}}{\underline{p_{i+} \sqrt{p_{+j}}}} \right) l_{jk} = \sum_{j=1}^n \left(\frac{q_{ij}}{\sqrt{p_{+j}}} \right) l_{jk} \quad (i \in I; k = 1, 2, \dots, K)$$

項目 J の第 j 選択肢の第 k 成分スコア

$$z_{jk}^* = \sum_{i=1}^m u_{ik} \underline{x_{ij}^*} = \sum_{i=1}^m \left(\frac{p_{ij}}{\underline{p_{+j} \sqrt{p_{i+}}}} \right) u_{ik} = \sum_{i=1}^m \left(\frac{q_{ij}^*}{\sqrt{p_{i+}}} \right) u_{ik} \quad (j \in J; k = 1, 2, \dots, K)$$

なぜ“この形”とするかが重要. これについてはあとで述べる.

添字, Σ の使い方に注意, 慣れる

「プロフィール」と「ストレッチ・プロフィール」

- ある行列つまり“データ”として, “プロフィール”と“ストレッチ・プロフィール”(重み付きプロフィール)を考えること.
- 行の質量 (p_{i+}), 列の質量 (p_{+j}) の正の平方根で“プロフィール”を重み調整している(ストレッチで伸ばす).

プロフィール

行側

$$q_{ij} = \frac{p_{ij}}{p_{i+}} = \frac{f_{ij}}{f_{i+}}$$

列側

$$q_{ij}^* = \frac{p_{ij}}{p_{+j}} = \frac{f_{ij}}{f_{+j}}$$

いわゆる相対確率
行和=1, 列和=1
とそろえたこと

ストレッチ・プロフィール



$$\begin{aligned} \frac{q_{ij}}{\sqrt{p_{+j}}} &= \frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} \quad (i \in I, j \in J) \\ \frac{q_{ij}^*}{\sqrt{p_{i+}}} &= \frac{p_{ij}}{p_{+j} \sqrt{p_{i+}}} \quad (i \in I, j \in J) \end{aligned}$$

この形に注目する
あとで図で観察する

プロフィールとは？

- いわゆる“相対確率”のこと.
- 日常的にクロス表の分析で行っていること.
- 行和 = 1 (100%)あるいは列和 = 1 (100%)とそろえること.
- “行プロフィール”と“列プロフィール”がある.
- 実際にこれらを例としたクロス表で求めてみる.

行のプロフィール

列のプロフィール

$$q_{ij} = \frac{p_{ij}}{p_{i+}} = \frac{f_{ij}}{f_{i+}}$$

$$q_{ij}^* = \frac{p_{ij}}{p_{+j}} = \frac{f_{ij}}{f_{+j}}$$

クロス表全体の観察

*住んでいる地区は緑(みどり)が多いと感じますか。

*いま住んでいるまちが気に入っていますか。

度数	1	2	3	4	5	合計
全体%						
列%						
行%						
1	166 8.53 54.43 31.68	239 12.28 27.19 45.61	86 4.42 18.49 16.41	26 1.34 11.40 4.96	7 0.36 10.14 1.34	524 26.93
2	131 6.73 42.95 10.61	598 30.73 68.03 48.42	324 16.65 69.68 26.23	146 7.50 64.04 11.82	36 1.85 52.17 2.91	1235 63.46
3	6 0.31 1.97 3.49	40 2.06 4.55 23.26	55 2.83 11.83 31.98	51 2.62 22.37 29.65	20 1.03 28.99 11.63	172 8.84
4	2 0.10 0.66 13.33	2 0.10 0.23 13.33	0 0.00 0.00 0.00	5 0.26 2.19 33.33	6 0.31 8.70 40.00	15 0.77
合計	305 15.67	879 45.17	465 23.90	228 11.72	69 3.55	1946

(行の質量)

(列の質量)

下のように対応する

$$\text{全体\%: } p_{ij} = \frac{f_{ij}}{N}$$

$$\text{行\%: } q_{ij} = \frac{p_{ij}}{p_{i+}} = \frac{f_{ij}}{f_{i+}} \Leftrightarrow \text{行プロフィール}$$

$$\text{列\%: } q_{ij}^* = \frac{p_{ij}}{p_{+j}} = \frac{f_{ij}}{f_{+j}} \Leftrightarrow \text{列プロフィール}$$

行の質量



列の比率の重心(平均比率)

列の質量



行の比率の重心(平均比率)

この意味はあとで調べる。

行プロファイルの観察

- “行プロファイル”の比率とそのグラフを表示, 観察する.
- 前ページのクロス表の諸量と対比し確認する.
- たとえば「1.大変気に入っている」を確認する.

<行プロファイル>

プロフィール

棒グラフ

カテゴリ名	1.かなり多い	2.多いほう	3.ふつう	4.少ないほう	5.少ない	周辺度数	周辺割合(%)
1.大変気に入っている						524	26.9
2.まあ気に入っている						1235	63.5
3.あまり気に入っていない						172	8.8
4.気に入っていない						15	0.8

表

カテゴリ名	1.かなり多い	2.多いほう	3.ふつう	4.少ないほう	5.少ない	周辺度数	周辺割合(%)
1.大変気に入っている	31.7	45.6	16.4	5.0	1.3	524	26.9
2.まあ気に入っている	10.6	48.4	26.2	11.8	2.9	1235	63.5
3.あまり気に入っていない	3.5	23.3	32.0	29.7	11.6	172	8.8
4.気に入っていない	13.3	13.3	0.0	33.3	40.0	15	0.8

(行の質量)

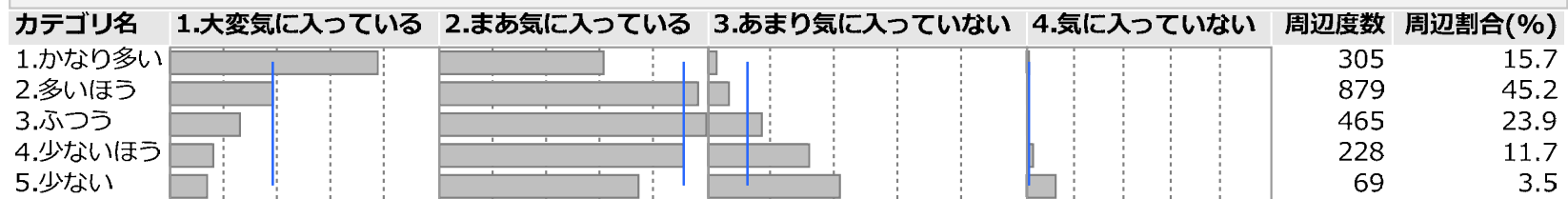
列プロフィールの観察

- “列プロフィール”の比率とそのグラフを表示，観察する.
- クロス表の諸量と対比し確認する.
- たとえば「1.かなり多い」を確認する.

<列プロフィール>

プロフィール

棒グラフ



表

カテゴリ名	1.大変気に入っている	2.まあ気に入っている	3.あまり気に入っていない	4.気に入っていない	周辺度数	周辺割合(%)
1.かなり多い	54.4	43.0	2.0	0.7	305	15.7
2.多いほう	27.2	68.0	4.6	0.2	879	45.2
3.ふつう	18.5	69.7	11.8	0.0	465	23.9
4.少ないほう	11.4	64.0	22.4	2.2	228	11.7
5.少ない	10.1	52.2	29.0	8.7	69	3.5

(列の質量) 48

ストレッチ・プロファイルとは？

- “プロファイル”を行の質量(p_{i+}), 列の質量(p_{+j})の正の平方根で重み調整(ストレッチ)し“ストレッチ・プロファイル”とする.
- この“分布”を“雲”(nuage)という.
- これは具体的に何を意味するのかが, 対応分析の特性を知る重要なポイント(後述する).

行のストレッチ・プロファイル

$$\frac{q_{ij}}{\sqrt{p_{+j}}} = \frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}} \quad (i \in I, j \in J)$$

列のストレッチ・プロファイル

$$\frac{q_{ij}^*}{\sqrt{p_{i+}}} = \frac{p_{ij}}{p_{+j}\sqrt{p_{i+}}} \quad (i \in I, j \in J)$$

前にみた“成分スコア”を再確認

- 求めた成分スコアの要約

		$k = 1$	$k = 2$	$k = 3$	$\mathbf{Z}_{m \times K}$
質問文	質問選択肢	第1成分スコア z_{i1}	第2成分スコア z_{i2}	第3成分スコア z_{i3}	
質問I: あなたは, いま住んでいるまちが気に入っていますか.	1. 大変気に入っている	-0.4442	0.2027	-0.0353	
	2. まあ気に入っている	0.0623	-0.1315	0.0311	
	3. あまり気に入っていない	0.7886	0.1907	-0.1698	
	4. 気に入っていない	1.3458	1.5567	0.6157	
質問文	質問選択肢	z_{j1}^*	z_{j2}^*	z_{j3}^*	
質問J: あなたの住んでいる地区は, 都市としては“緑(みどり)が多い”と感じますか.	1. かなり多い	-0.5403	0.3235	-0.0640	
	2. 多いほう	-0.1118	-0.1050	0.0000	$\mathbf{Z} = \begin{pmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ z_{31} & z_{32} & z_{33} \\ z_{41} & z_{42} & z_{43} \end{pmatrix}$
	3. ふつう	0.1545	-0.1500	0.0000	
	4. 少ないほう	0.5530	0.0750	0.0000	
	5. 少ない	0.9438	0.6800	0.0000	

ここは行成分スコア

行成分スコアと列成分スコア

- 以下のように対応する.

行成分スコア

$$\mathbf{Z}_{4 \times 3} = (z_{ik}) = \begin{matrix} & \begin{matrix} k=1 & k=2 & k=3 \end{matrix} \\ \begin{matrix} i=1 \\ i=2 \\ i=3 \\ i=4 \end{matrix} & \begin{pmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \\ z_{31} & z_{32} & z_{33} \\ z_{41} & z_{42} & z_{43} \end{pmatrix} \end{matrix}$$

$$z_{ik} = \sum_{j=1}^n \left(\frac{q_{ij}}{\sqrt{p_{+j}}} \right) l_{jk} = \sum_{j=1}^n \left(\frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} \right) l_{jk}$$

列成分スコア

$$\mathbf{Z}^*_{5 \times 3} = (z^*_{jk}) = \begin{matrix} & \begin{matrix} k=1 & k=2 & k=3 \end{matrix} \\ \begin{matrix} j=1 \\ j=2 \\ j=3 \\ j=4 \\ j=5 \end{matrix} & \begin{pmatrix} * & * & * \\ z^*_{11} & z^*_{12} & z^*_{13} \\ * & * & * \\ z^*_{21} & z^*_{22} & z^*_{23} \\ * & * & * \\ z^*_{31} & z^*_{32} & z^*_{33} \\ * & * & * \\ z^*_{41} & z^*_{42} & z^*_{43} \\ * & * & * \\ z^*_{51} & z^*_{52} & z^*_{53} \end{pmatrix} \end{matrix}$$

$$z^*_{jk} = \sum_{i=1}^m \left(\frac{q^*_{ij}}{\sqrt{p_{i+}}} \right) u_{ik} = \sum_{i=1}^m \left(\frac{p_{ij}}{p_{+j} \sqrt{p_{i+}}} \right) u_{ik}$$

“合成変数”の形を確認

- 成分式は、以下のように“ストレッチ・プロファイルの加重和”の形になっている。
- “合成変数”とはこうした加重和で得られるスコアのこと。

行成分スコア

$$z_{ik} = \sum_{j=1}^n \left(\frac{q_{ij}}{\sqrt{p_{+j}}} \right) l_{jk} \Leftrightarrow l_{jk} \text{を係数とする} \left(\frac{q_{ij}}{\sqrt{p_{+j}}} \right) \text{の和}$$

この”係数“は何か
(固有ベクトル, 特異ベクトルの要素)

これは何か(★)
ストレッチ・プロファイル

列成分スコア

$$z_{jk}^* = \sum_{i=1}^m \left(\frac{q_{ij}^*}{\sqrt{p_{i+}}} \right) u_{ik} \Leftrightarrow u_{ik} \text{を係数とする} \left(\frac{q_{ij}^*}{\sqrt{p_{i+}}} \right) \text{の和}$$

ストレッチ・プロフィールを要素とする行列

- 行または列のストレッチ・プロフィールを要素とする行列を考える. (←前に“ある行列”としたもの)
- この要素をデータ(x_{ij})とする分散共分散行列を作り固有値問題として解く, あるいは特異値分解を行う.
- 行側から解く, 列側から解く, があるがどちらから行っても解は同じ(対称の対応分析のとき).

行ストレッチ・プロフィール

$$x_{ij} = \frac{q_{ij}}{\sqrt{p_{+j}}} = \frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}} \quad (i \in I, j \in J)$$

⇓

$$\mathbf{X}_{m \times n} = (x_{ij}) = \begin{pmatrix} \frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}} \end{pmatrix} = \begin{pmatrix} \frac{q_{ij}}{\sqrt{p_{+j}}} \end{pmatrix}$$

列ストレッチ・プロフィール

$$\frac{q_{ij}^*}{\sqrt{p_{i+}}} = \frac{p_{ij}}{p_{+j}\sqrt{p_{i+}}} \quad (i \in I, j \in J)$$

⇓

$$\mathbf{X}^+_{m \times n} = (x_{ij}^*) = \begin{pmatrix} \frac{p_{ij}}{p_{+j}\sqrt{p_{i+}}} \end{pmatrix} = \begin{pmatrix} \frac{q_{ij}^*}{\sqrt{p_{i+}}} \end{pmatrix}$$

★メモ：データ行列は何か？

- 数理的，形式的には，このストレッチ・プロファイルを要素とする行列を“出発行列”として，この特異値分解あるいは共分散行列の固有値問題を解く．
- 出発行列の作り方はいくつかあるが解は同じ．
- テキスト，第Ⅱ部，19～20ページあたりに要約．

$$\mathbf{X}_{m \times n} = (x_{ij}) = \left(\frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} \right) \left(\begin{array}{l} \text{行ストレッチ・プロファイル} \\ \text{を要素とする行列} \end{array} \right)$$

$$\mathbf{Q} = (y_{ij}) = \left(\frac{p_{ij}}{\sqrt{p_{i+} p_{+j}}} \right) = \left(\frac{f_{ij}}{\sqrt{f_{i+} f_{+j}}} \right) \quad (i \in I, j \in J)$$

↓ \mathbf{X} のかわりにこれを用いても解は同じ．
 $m \times n$

$$\mathbf{V}_{n \times n}^* = \mathbf{Q}^t \mathbf{Q} = \mathbf{P}_J^{-1/2} \mathbf{P}_{JI} \mathbf{P}_I^{-1} \mathbf{P}_{IJ} \mathbf{P}_J^{-1/2} \quad (\text{この行列，共分散行列の固有値問題})$$

$$\text{tr}(\mathbf{V}^*) - 1 = \sum_{k=1}^K \lambda_k = \frac{\chi_p^2}{N} \quad (K = \min\{m, n\} - 1) \text{ (こうなる)}$$

たとえば、行の“第1成分スコア”について

$$z_{ik} = \sum_{j=1}^n l_{jk} x_{ij} = \sum_{j=1}^n \left(\frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} \right) l_{jk} = \sum_{j=1}^n \left(\frac{q_{ij}}{\sqrt{p_{+j}}} \right) l_{jk} \quad (i \in I; k = 1, 2, \dots, K)$$

たとえば、 $k=1$ とし、行側 $i \in I$ (4つの選択肢)のスコア z_{i1} を示す。
行選択肢 $i \in I$ の第1成分スコアという“合成変数”は以下。

$$z_{i1} = \sum_{j=1}^5 l_{j1} x_{ij} = \underline{l_{11} x_{i1}} + \underline{l_{21} x_{i2}} + \underline{l_{31} x_{i3}} + \underline{l_{41} x_{i4}} + \underline{l_{51} x_{i5}} \quad \textcircled{\text{多次元(多変量)が“1次元”化された}}$$

$$= \underbrace{\frac{p_{i1}}{p_{i+} \sqrt{p_{+1}}}}_{\text{「1. かなり多い」}} \underline{l_{11}} + \underbrace{\frac{p_{i2}}{p_{i+} \sqrt{p_{+2}}}}_{\text{「2. 多いほう」}} \underline{l_{21}} + \underbrace{\frac{p_{i3}}{p_{i+} \sqrt{p_{+3}}}}_{\text{「3. ふつう」}} \underline{l_{31}} + \underbrace{\frac{p_{i4}}{p_{i+} \sqrt{p_{+4}}}}_{\text{「4. 少ないほう」}} \underline{l_{41}} + \underbrace{\frac{p_{i5}}{p_{i+} \sqrt{p_{+5}}}}_{\text{「5. 少ない」}} \underline{l_{51}}$$

列側(質問項目 J)の5つの選択肢の“合成変数”つまり加重和、
重み付き平均を作った!!!

“係数”は、いわゆる固有ベクトル・特異ベクトルの要素(後述)

(つづき)

- ここで、(行の側の)第1成分という“1次元”(の合成変数)の情報に、列の項目 J の5つの「選択肢」が“5変数情報”の加重和として入っている.
- 第2成分スコア z_{i2} についても同様に $k=2$ とした5変数の加重和がある.
- この例では成分数として、あわせて、 $K=\min\{m,n\}-1=3$ つの成分がある(第1成分～第3成分).
- つまり、3つの“合成変数”があるということ.

(つづき)

- いま, 行の項目 I の4つの選択肢すべてに対する“第1成分スコア”を求める式を書き下してみる. 次ページ.
- これは列の側の成分スコアは, 項目 I の4つの選択肢のストレッチ・プロファイルの加重和となる.
- 重みとなる“係数”は成分により異なる(固有値に対応して変わる). 加重で影響度を調整している.
- 合成変数の重み(加重)として, 固有ベクトルあるいは特異値ベクトルの要素を使う(数理的にそうなる).

(つづき)

行の質問項目 I の選択肢 i の 第1成分スコア すべてを求める.

$$z_{i1} = \underbrace{\frac{p_{i1}}{p_{i+}\sqrt{p_{+1}}}}_{\text{「1. かなり多い」}} l_{11} + \underbrace{\frac{p_{i2}}{p_{i+}\sqrt{p_{+2}}}}_{\text{「2. 多いほう」}} l_{21} + \underbrace{\frac{p_{i3}}{p_{i+}\sqrt{p_{+3}}}}_{\text{「3. ふつう」}} l_{31} + \underbrace{\frac{p_{i4}}{p_{i+}\sqrt{p_{+4}}}}_{\text{「4. 少ないほう」}} l_{41} + \underbrace{\frac{p_{i5}}{p_{i+}\sqrt{p_{+5}}}}_{\text{「5. 少ない」}} l_{51}$$

行の第1成分スコア

$$z_{11} = \frac{p_{11}}{p_{1+}\sqrt{p_{+1}}} l_{11} + \frac{p_{12}}{p_{1+}\sqrt{p_{+2}}} l_{21} + \frac{p_{13}}{p_{1+}\sqrt{p_{+3}}} l_{31} + \frac{p_{14}}{p_{1+}\sqrt{p_{+4}}} l_{41} + \frac{p_{15}}{p_{1+}\sqrt{p_{+5}}} l_{51}$$

1. 大変気に入っている

$$z_{21} = \frac{p_{21}}{p_{2+}\sqrt{p_{+1}}} l_{11} + \frac{p_{22}}{p_{2+}\sqrt{p_{+2}}} l_{21} + \frac{p_{23}}{p_{2+}\sqrt{p_{+3}}} l_{31} + \frac{p_{24}}{p_{2+}\sqrt{p_{+4}}} l_{41} + \frac{p_{25}}{p_{2+}\sqrt{p_{+5}}} l_{51}$$

2. まあ気に入っている

$$z_{31} = \frac{p_{31}}{p_{3+}\sqrt{p_{+1}}} l_{11} + \frac{p_{32}}{p_{3+}\sqrt{p_{+2}}} l_{21} + \frac{p_{33}}{p_{3+}\sqrt{p_{+3}}} l_{31} + \frac{p_{34}}{p_{3+}\sqrt{p_{+4}}} l_{41} + \frac{p_{35}}{p_{3+}\sqrt{p_{+5}}} l_{51}$$

3. あまり気に入っていない

$$z_{41} = \frac{p_{41}}{p_{4+}\sqrt{p_{+1}}} l_{11} + \frac{p_{42}}{p_{4+}\sqrt{p_{+2}}} l_{21} + \frac{p_{43}}{p_{4+}\sqrt{p_{+3}}} l_{31} + \frac{p_{44}}{p_{4+}\sqrt{p_{+4}}} l_{41} + \frac{p_{45}}{p_{4+}\sqrt{p_{+5}}} l_{51}$$

4. 気に入っていない

「1」に対応している

$$i \in I = \{1, 2, 3, 4\}$$

★補足メモ: 別のデータ表とその要素

- ここから71ページまでは「参考情報」とする.
- ストレッチ・プロファイルを要素とするデータ行列とは異なるデータ行列を作ることと, 成分スコアの算出を説明.
- カイ二乗統計量の式にその要素は含まれる.
- これはストレッチ・プロファイルからのデータ行列と同じ結果となることが知られている.

$$\begin{aligned}\chi_p^2 &= \sum_{i=1}^m \sum_{j=1}^n \frac{N(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}} = \sum_{i=1}^m \sum_{j=1}^n \frac{\left(f_{ij} - \frac{f_{i+}f_{+j}}{N}\right)^2}{\frac{f_{i+}f_{+j}}{N}} \\ &= \sum_{i=1}^m \sum_{j=1}^n \frac{(f_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^m \sum_{j=1}^n \left(\frac{f_{ij} - e_{ij}}{\sqrt{e_{ij}}} \right)^2\end{aligned}$$

★カイ二乗統計量と総変動の関係

$$\frac{\chi_p^2}{N} = \sum_{i=1}^m \sum_{j=1}^n \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}} = \sum_{i=1}^m \sum_{j=1}^n \left(\frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}} \right)^2 = \sum_{k=1}^K \lambda_k$$

総変動(全慣性)
=固有値の総和

つぎを要素とするデータ行列の“特異値分解”を考えることに同じこと. テキスト参照. これも“ある行列”の1つ.

$$\underline{y_{ij}^*} = \frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}} \left(= \frac{f_{ij} - e_{ij}}{\sqrt{e_{ij}}} \right) \Rightarrow \mathbf{Y}_{m \times n}^* = (y_{ij}^*)$$

(y_{ij}^* を要素とする行列)

★さらに, ...

$$y_{ij}^* = \frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}} \Rightarrow \mathbf{Y}_{m \times n}^* = \left(y_{ij}^* \right) \quad (y_{ij}^* \text{を要素とする行列})$$

$\left(\begin{array}{l} \text{もとの } f_{ij}, f_{i+}, f_{+j} \text{ あるいは } p_{ij}, p_{i+}, p_{+j} \text{ から, あらたに,} \\ y_{ij}^* \text{ を作る. } \underline{\text{これを要素とするあらたなデータ表とする.}} \end{array} \right)$

$$\mathbf{Y}^* = \left(y_{ij}^* \right) = \mathbf{P}_I^{-1/2} \left(\mathbf{P}_{IJ} - \mathbf{r} \mathbf{c}^t \right) \mathbf{P}_J^{-1/2}$$

$$\mathbf{X}_{m \times n} = \left(x_{ij} \right) = \left(\frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} \right) = \left(\frac{q_{ij}}{\sqrt{p_{+j}}} \right)$$

行列, ベクトルで表す.
 II 部, 14ページあたり
 どちらでもよい.

- なぜ, ストレッチ・プロファイルやこうしたことを行うのか?
- これはうしろに数値例を用いて, もう一度述べる.
- ここは特異値分解と, 得られる特異ベクトル(つまり合成変数)の係数)を数値例で調べる.

★“特異値分解”で行列を分解

$$\mathbf{Y}^* = \begin{pmatrix} y_{ij}^* \end{pmatrix} = \mathbf{P}_I^{-1/2} \left(\mathbf{P}_{IJ} - \mathbf{r}\mathbf{c}^t \right) \mathbf{P}_J^{-1/2}$$

⇕

$$\mathbf{Y}^*_{m \times n} = \underbrace{\mathbf{U}}_{m \times K} \underbrace{\mathbf{\Lambda}^{1/2}}_{K \times K} \underbrace{\mathbf{L}^t}_{K \times n}$$

$\left(\begin{array}{l} \mathbf{L}^t \text{ は } \mathbf{L} \text{ の転置行列} \\ \mathbf{\Lambda}^{1/2} = \text{diag} \left(\sqrt{\lambda_k} \right) = \text{diag} \left(\alpha_k \right) \end{array} \right)$

この行列を



下のこの行列のように
“分解”する操作

$$\mathbf{U}_{m \times K} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k, \dots, \mathbf{u}_K)$$

$$\mathbf{u}_k^t = (u_{1k}, u_{2k}, \dots, u_{ik}, \dots, u_{mk})$$

左特異ベクトルとその行列

$$\mathbf{L}_{n \times K} = (\mathbf{l}_1, \mathbf{l}_2, \dots, \mathbf{l}_k, \dots, \mathbf{l}_K)$$

$$\mathbf{l}_k^t = (l_{1k}, l_{2k}, \dots, l_{jk}, \dots, l_{nk})$$

右特異ベクトルとその行列

これが、データ要素(ストレッチ・プロファイル)の“**係数**”
となり成分スコア(合成変数)がえられる。

特異値分解: ある行列を上のように3つの行列に分けること。

★「行」側(質問I)からみる

$$\mathbf{z}_k = \underbrace{\mathbf{P}_I^{-1} \mathbf{P}_{IJ} \mathbf{P}_J^{-1/2}}_{m \times n} \mathbf{l}_k \quad (k = 1, 2, \dots, K)$$

$$\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k, \dots, \mathbf{z}_K) = \underbrace{\mathbf{P}_I^{-1} \mathbf{P}_{IJ} \mathbf{P}_J^{-1/2}}_{m \times n} \mathbf{L} = \mathbf{X} \mathbf{L}$$

項目Iの第i選択肢の第k成分スコア

上の行列の(i,k)要素

$$z_{ik} = \sum_{j=1}^n l_{jk} x_{ij} = \sum_{j=1}^n \left(\frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} \right) l_{jk} = \sum_{j=1}^n \left(\frac{q_{ij}}{\sqrt{p_{+j}}} \right) l_{jk} \quad (i \in I; k = 1, 2, \dots, K)$$

ここで, $\mathbf{X} = (x_{ij}) = \left(\frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} \right) = \left(\frac{q_{ij}}{\sqrt{p_{+j}}} \right)$, これがデータ

プロフィールを質量の平方根でストレッチ, “ストレッチ・プロフィール”
なぜこうするか, どういう意味があるのか

★「列」側(質問 J)からみる

$$\mathbf{z}_k^* = \underbrace{\mathbf{P}_J^{-1} \mathbf{P}_{JI} \mathbf{P}_I^{-1/2}}_{n \times m} \mathbf{u}_k \quad (k = 1, 2, \dots, K)$$

$$\mathbf{Z}^* = \left(\mathbf{z}_1^*, \mathbf{z}_2^*, \dots, \mathbf{z}_k^*, \dots, \mathbf{z}_K^* \right) \underbrace{\mathbf{P}_J^{-1} \mathbf{P}_{JI} \mathbf{P}_I^{-1/2}}_{n \times m} \mathbf{U} = \left(\mathbf{X}^* \right)^t \mathbf{U}$$

上の行列の (j, k) 要素

項目 J の第 j 選択肢の第 k 成分スコア

$$z_{jk}^* = \sum_{i=1}^m u_{ik} x_{ij}^* = \sum_{i=1}^m \left(\frac{p_{ij}}{p_{+j} \sqrt{p_{i+}}} \right) u_{ik} = \sum_{i=1}^m \left(\frac{q_{ij}^*}{\sqrt{p_{i+}}} \right) u_{ik} \quad (j \in J; k = 1, 2, \dots, K)$$

ここで, $\mathbf{X}^* = (x_{ij}^*) = \left(\frac{p_{ij}}{p_{+j} \sqrt{p_{i+}}} \right) = \left(\frac{q_{ij}^*}{\sqrt{p_{i+}}} \right)$, これがデータ

(ストレッチ・プロフィールが要素)

★数値例で確認(簡単に)

- ここで, “何を求めたか”を数値例で確認する.
- こうした操作はコンピュータ内で, それぞれの計算に適した“アルゴリズム”で処理される.
- “こんなことが行われる”ということを知っておけばよい.
- 通常, 処理アルゴリズムは, ここでみた式のフォローのようには行っていないことに注意する.
- 数値計算法, アルゴリズムそしてそれを実現するプログラムがあること. (“近似計算”となることもある)
- とくに, 寸法の大きな2元データ表の処理は, アルゴリズム上の工夫が必要となる. 例: **テキスト型データ**
- 対応分析法の利用上は, 各値が“何を意味するか”を知ること(意味, 解釈).

数値で確かめる(1)

- 固有ベクトルとその行列(合成変数の重み)
- 固有値・特異値から対角行列を作る

$$\mathbf{U}_{m \times K} = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k, \dots, \mathbf{u}_K) \Rightarrow \mathbf{U}_{4 \times 3} = \begin{pmatrix} -0.6532 & 0.50186 & -0.2285 \\ 0.14064 & -0.4997 & 0.30961 \\ 0.6644 & 0.27055 & -0.6301 \\ 0.33484 & 0.65208 & 0.67451 \end{pmatrix}$$

左特異ベクトルとその行列
これが列成分スコアの重み(係数)であった

$$\mathbf{L}_{n \times K} = (\mathbf{l}_1, \mathbf{l}_2, \dots, \mathbf{l}_k, \dots, \mathbf{l}_K) \Rightarrow \mathbf{L}_{5 \times 3} = \begin{pmatrix} -0.6062 & 0.61114 & -0.3162 \\ -0.2129 & -0.3383 & 0.55127 \\ 0.21405 & -0.3512 & -0.3739 \\ 0.53643 & 0.12254 & -0.4566 \\ 0.50362 & 0.61134 & 0.49782 \end{pmatrix}$$

右特異ベクトルとその行列
これが行成分スコアの重み(係数)であった

$$\Lambda_{K \times K}^{1/2} = \text{diag}(\sqrt{\lambda_k}) = \text{diag}(\alpha_k) \Rightarrow \Lambda_{3 \times 3}^{1/2} = \begin{pmatrix} 0.35288 & 0 & 0 \\ 0 & 0.20959 & 0 \\ 0 & 0 & 0.08012 \end{pmatrix}$$

特異値を対角要素とする対角行列

数値で確かめる(2)

- 行列の積を作り特異値分解を確かめる.

$$\begin{array}{c}
 \mathbf{Y}^* = \mathbf{U} \mathbf{\Lambda}^{1/2} \mathbf{L}^t \\
 \begin{matrix} m \times n & m \times K & K \times K & K \times n \end{matrix}
 \end{array}
 = \begin{pmatrix} -0.6532 & 0.50186 & -0.2285 \\ 0.14064 & -0.4997 & 0.30961 \\ 0.6644 & 0.27055 & -0.6301 \\ 0.33484 & 0.65208 & 0.67451 \end{pmatrix} \times \begin{pmatrix} 0.35288 & 0 & 0 \\ 0 & 0.20959 & 0 \\ 0 & 0 & 0.08012 \end{pmatrix} \times \begin{pmatrix} -0.6062 & -0.2129 & 0.21405 & 0.53643 & 0.50362 \\ 0.61114 & -0.3383 & -0.3512 & 0.12254 & 0.61134 \\ -0.3162 & 0.55127 & -0.3739 & -0.4566 & 0.49782 \end{pmatrix}$$

左特異ベクトルとその行列

×

特異値を対角要素とする行列

×

右特異ベクトルとその行列の転置

$$\begin{array}{c}
 \mathbf{Y}^* = \mathbf{U} \mathbf{\Lambda}^{1/2} \mathbf{L}^t \\
 \begin{matrix} m \times n & m \times K & K \times K & K \times n \end{matrix}
 \end{array}
 \quad \text{(このY*, つまり次ページY*を上のように分解した)}$$

$$= \begin{pmatrix} 0.209801 & 0.003395 & -0.079434 & -0.102396 & -0.060892 \\ -0.101936 & 0.038541 & 0.038130 & 0.002461 & -0.026685 \\ -0.091504 & -0.096930 & 0.049146 & 0.155767 & 0.127607 \\ -0.005189 & -0.041598 & -0.042916 & 0.055453 & 0.169964 \end{pmatrix} \quad (\star)$$

数値で確かめる(3)

- 前に調べた行列 \mathbf{Y}^* を行列, ベクトルで書き替える.

$$\mathbf{Y}_{m \times n}^* = \left(y_{ij}^* \right) = \left(\frac{p_{ij} - p_{i+} p_{+j}}{\sqrt{p_{i+} p_{+j}}} \right)$$

$$\mathbf{Y}^* = \left(y_{ij}^* \right) = \mathbf{P}_I^{-1/2} \left(\mathbf{P}_{IJ} - \mathbf{r} \mathbf{c}^t \right) \mathbf{P}_J^{-1/2}$$

$$= \begin{pmatrix} 0.209799 & 0.003406 & -0.079435 & -0.102398 & -0.060898 \\ -0.101939 & 0.038541 & 0.038129 & 0.002456 & -0.026685 \\ -0.091502 & -0.096936 & 0.049151 & 0.155774 & 0.127605 \\ -0.005189 & -0.041588 & -0.042917 & 0.055446 & 0.169969 \end{pmatrix}$$

$$\mathbf{r}_{m \times 1} = \begin{pmatrix} p_{1+} \\ p_{2+} \\ \vdots \\ p_{i+} \\ \vdots \\ p_{m+} \end{pmatrix}, \mathbf{c}_{n \times 1} = \begin{pmatrix} p_{+1} \\ p_{+2} \\ \vdots \\ p_{+j} \\ \vdots \\ p_{+n} \end{pmatrix}$$

行の“質量” 列の“質量”

(★★)

(★)と(★★)は等しい(わずかの計算誤差がある)
 ここは, 数値の確認のみ. 質量については後述.

★補足:

- “固有値問題”で, 固有ベクトルとの関係を示すと以下のように対応している.
- つまり, 特異値分解とは同じことを言い換えている.

$\mathbf{Y}_{m \times n}^* (\mathbf{Y}_{n \times m}^*)^t$ の固有値問題を解いて得られる固有ベクトル行列 \Leftrightarrow 行列 $\mathbf{U}_{m \times K}$

$(\mathbf{Y}_{n \times m}^*)^t \mathbf{Y}_{m \times n}^*$ の固有値問題を解いて得られる固有ベクトル行列 \Leftrightarrow 行列 $\mathbf{L}_{n \times K}$

$$\text{tr} \left[\mathbf{Y}_{m \times n}^* (\mathbf{Y}_{n \times m}^*)^t \right] = \text{tr} \left[(\mathbf{Y}_{n \times m}^*)^t \mathbf{Y}_{m \times n}^* \right] = \frac{\chi_p^2}{N} = \sum_{k=1}^K \lambda_k \left(K = \min \{m, n\} - 1 \right)$$

数値で確認：“行成分スコア”を求めてみる

$$\mathbf{X}_{m \times n} = (x_{ij}) = \left(\frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} \right) = \left(\frac{q_{ij}}{\sqrt{p_{+j}}} \right)$$

この“ストレッチ・プロファイル”を要素とするデータ表を考える
その理由はあとで(もう一度)述べる

クロス表(F表)から作った
あらたなデータ

$$\mathbf{X}_{m \times n} = (x_{ij}) = \left(\frac{q_{ij}}{\sqrt{p_{+j}}} \right) \Rightarrow \mathbf{X}_{4 \times 5} = \begin{pmatrix} 0.8002 & 0.6786 & 0.3357 & 0.1450 & 0.0709 \\ 0.2679 & 0.7205 & 0.5367 & 0.3454 & 0.1548 \\ 0.0881 & 0.3460 & 0.6542 & 0.8663 & 0.6175 \\ 0.3368 & 0.1984 & 0.0000 & 0.9738 & 2.1243 \end{pmatrix}$$

$$\mathbf{L}_{n \times K} \Rightarrow \mathbf{L}_{5 \times 3} = \begin{pmatrix} \lambda_1 & \lambda_2 & \lambda_3 \\ -0.6062 & 0.61114 & -0.3162 \\ -0.2129 & -0.3383 & 0.55127 \\ 0.21405 & -0.3512 & -0.3739 \\ 0.53643 & 0.12254 & -0.4566 \\ 0.50362 & 0.61134 & 0.49782 \end{pmatrix}$$

第1成分の加重

第2成分の加重

第3成分の加重

固有ベクトル行列, これが“係数”
加重和を作る, つまり“合成変数”

固有ベクトル ⇔ 特異ベクトル (言い換えただけ)

$$z_{ik} = \sum_{j=1}^n \left(\frac{q_{ij}}{\sqrt{p_{+j}}} \right) l_{jk}$$

スライド, 66ページ

(つづき)

$$\mathbf{Z}_{m \times K} = \underbrace{\mathbf{P}_I^{-1} \mathbf{P}_{IJ} \mathbf{P}_J^{-1/2}}_{m \times n} \mathbf{L}_{n \times K} = \mathbf{X}_{m \times n} \mathbf{L}_{n \times K} \Rightarrow \mathbf{X}_{4 \times 5} \mathbf{L}_{5 \times 3} =$$

λ_1	λ_2	λ_3
\Downarrow	\Downarrow	\Downarrow
$k=1$	$k=2$	$k=3$
z_{11}	z_{12}	z_{13}
z_{21}	z_{22}	z_{23}
z_{31}	z_{32}	z_{33}
z_{41}	z_{42}	z_{43}

Z
 $m \times K$

z_{i1}	z_{i2}	z_{i3}
-0.4442	0.2027	-0.0353
0.0623	-0.1315	0.0311
0.7886	0.1907	-0.1698
1.3458	1.5567	0.6157

- 1.大変気に入ってる
- 2.まあ気に入ってる
- 3.あまり気に入ってない
- 4.気に入ってない

Z*
 $n \times K$

「行成分スコア」
当たり前だが前の表に一致する
“列成分スコア”も同じように得られる 確認



ここからの目標

- 対応分析の基本要素を，数値例と若干の数式を用いて“形式的に”説明した，記法，符丁を確認のうえ次に進む．
- ここからは，別の数値例を用いて，より具体的に対応分析の仕組みを調べる．
 - 2元データ表の要件の再確認
 - プロファイルとストレッチ・プロファイルの“幾何学特性”
 - カイ二乗距離とユークリッド距離
 - 総変動(全慣性)と距離の関係
 - 総変動(全慣性)と固有値ほかの関係
 - 寄与度(絶対寄与度，相対寄与度・平方相関)
- 最後に数値例で確認する．

★再確認: 2元データ表とは？

- 対応分析法で扱う“2元データ表”の条件を要約する.
- 典型的な例が“2値データ表”(インシデント行列)や“2元クロス表”(あるいは分割表)である.
- これをさらに緩めて, いろいろなデータ表に対応分析法を利用できる.
- このことが, 自由回答・自由記述に代表されるテキスト型データ(textual data)の分析で威力を発揮する.

例1: 回答者 × 抽出語句 (寸法: 314s × 139w)

例2: 抽出語句 × 性年齢区分 (寸法: 139w × 15区分)

(*) テキスト, 第 I 部, 52ページ(5.2.1節の例), インターネット使い方
“「プラス」になるとおもうこと”

2元データ表の要件

- ① 原則, “2元”の行列形式であること.
 - ② データ表の各要素(各セル内の値)が“非負の数値”であること.
 - ③ 行あるいは列の比率(割合)パターン, “プロフィール が意味のある”データ.
 - ④ あるいはそれに相当する場面を想定できる2元データ表.
 - ⑤ “(2元)クロス表”元に主な記法, とくにプロフィール, ストレッチ・プロフィール, 質量, 重心などの表記.
 - ⑥ これを一般の“2元データ表”と読み替えて考える.
- 以上を再確認し, かつ追加情報を述べる.

★2元クロス表(2元データ表の典型例)

- ここから82ページまでは, 符丁を確認するだけ.
- 2元データ表の典型例として“2元クロス表”で示す.
- トイ・データを用いてさらに説明する.

$$\mathbf{F}_{m \times n} = (f_{ij}) \left(\begin{array}{l} f_{ij} \geq 0 \\ i \in I, j \in J \end{array} \right)$$

寸法が $m \times n$ のクロス表

$$I = \{1, 2, \dots, i, \dots, m\}$$

$$J = \{1, 2, \dots, j, \dots, n\}$$

項目 I の m 個の選択肢と
項目 J の n 個の選択肢

“選択肢”は, カテゴリー, 分類区分のこと
対応分析法の言い方では“フォルム”(forme)あるいは“モダリテ”(modalité)という

★2元クロス表 $F=(f_{ij})$ の構成

- 2元クロス表の度数(頻度)分布を以下のように表す.

$I \backslash J$	1	2	...	j	...	n	行和
1	f_{11}	f_{12}	...	f_{1j}	...	f_{1n}	f_{1+}
2	f_{21}	f_{22}	...	f_{2j}	...	f_{2n}	f_{2+}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
i	f_{i1}	f_{i2}	...	f_{ij}	...	f_{in}	f_{i+}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
m	f_{m1}	f_{m2}	...	f_{mj}	...	f_{mn}	f_{m+}
列和	f_{+1}	f_{+2}	...	f_{+j}	...	f_{+n}	f_{++}

項目 I の周辺分布

$$f_{i+} = \sum_{j=1}^n f_{ij}$$

$\mathbf{F} = (f_{ij})_{m \times n}$

項目 I, J の同時分布

総度数

$$f_{++} = \sum_{i=1}^m \sum_{j=1}^n f_{ij} = N$$

項目 J の周辺分布

$$f_{+j} = \sum_{i=1}^m f_{ij}$$

★記号・記法の用意・準備

(i, j)セル内の度数: f_{ij} ($i \in I, j \in J$)

行和: $f_{i+} = \sum_{j=1}^n f_{ij}$

列和: $f_{+j} = \sum_{i=1}^m f_{ij}$

「+」和記号を使ったこれらの記法に慣れよう！

総度数: $\sum_{i=1}^m f_{i+} = \sum_{j=1}^n f_{+j} = N$

$$f_{++} = \sum_{i=1}^m \sum_{j=1}^n f_{ij} = N$$

★これを確率行列 $\mathbf{P}=(p_{ij})$ にする

$$p_{ij} = \frac{f_{ij}}{N} \quad (\text{セル}(i, j) \text{の同時確率})$$

$$p_{i+} = \frac{f_{i+}}{N} = \frac{\sum_{j=1}^n f_{ij}}{N} \quad (\text{列のセル}(i, +) \text{の周辺確率})$$

$$p_{+j} = \frac{f_{+j}}{N} = \frac{\sum_{i=1}^m f_{ij}}{N} \quad (\text{列のセル}(+, j) \text{の周辺確率})$$

実現確率
(相対確率)

$$\sum_{i=1}^m \sum_{j=1}^n p_{ij} = \sum_{i=1}^m p_{i+} = \sum_{j=1}^n p_{+j} = 1 \quad (\text{全確率} = 1)$$

これをあらためて“確率行列” $\mathbf{P}=(p_{ij})$ にする.

★クロス表から作った“確率行列” $\mathbf{P}=(p_{ij})$ の構成

$I \backslash J$	1	2	...	j	...	n	行和	列の平均 プロフィール
1	p_{11}	p_{12}	...	p_{1j}	...	p_{1n}	p_{1+}	$\mathbf{r}_{m \times 1} = \begin{pmatrix} p_{1+} \\ p_{2+} \\ \vdots \\ p_{i+} \\ \vdots \\ p_{m+} \end{pmatrix}$
2	p_{21}	p_{22}	...	p_{2j}	...	p_{2n}	p_{2+}	
\vdots	\vdots	\vdots	$\mathbf{P}_{IJ} = \frac{1}{N} \mathbf{F} = (p_{ij})$			\vdots	\vdots	
i	p_{i1}	p_{i2}	...	p_{ij}	...	p_{in}	p_{i+}	
\vdots	\vdots	\vdots		\vdots		\vdots	\vdots	
m	p_{m1}	p_{m2}	...	p_{mj}	...	p_{mn}	p_{m+}	
列和	p_{+1}	p_{+2}	...	p_{+j}	...	p_{+n}	1	行の“質量”
行の平均 プロフィール	$\mathbf{c}_{1 \times n} = (p_{+1}, p_{+2}, \dots, p_{+j}, \dots, p_{+n})^t$							質量(mass)

列の“質量”

★さらに以下を用意(行列で表記しただけ)

$$\mathbf{P}_{IJ} = \frac{1}{N} \mathbf{F} = (p_{ij}) \quad (i \in I, j \in J) \quad (\text{同時確率分布})$$

$$\mathbf{P}_I = \text{diag}(p_{i+}) \quad (i \in I) \quad (\text{行の周辺確率分布; 行質量の分布})$$

$$\mathbf{P}_J = \text{diag}(p_{+j}) \quad (j \in J) \quad (\text{列の周辺確率分布; 列質量の分布})$$

$$p_{i+} = \sum_{j=1}^n p_{ij}, \quad p_{+j} = \sum_{i=1}^m p_{ij}$$

$$\sum_{i=1}^m p_{i+} = \sum_{j=1}^n p_{+j} = 1$$

$$\sum_{i=1}^m \sum_{j=1}^n p_{ij} = 1$$

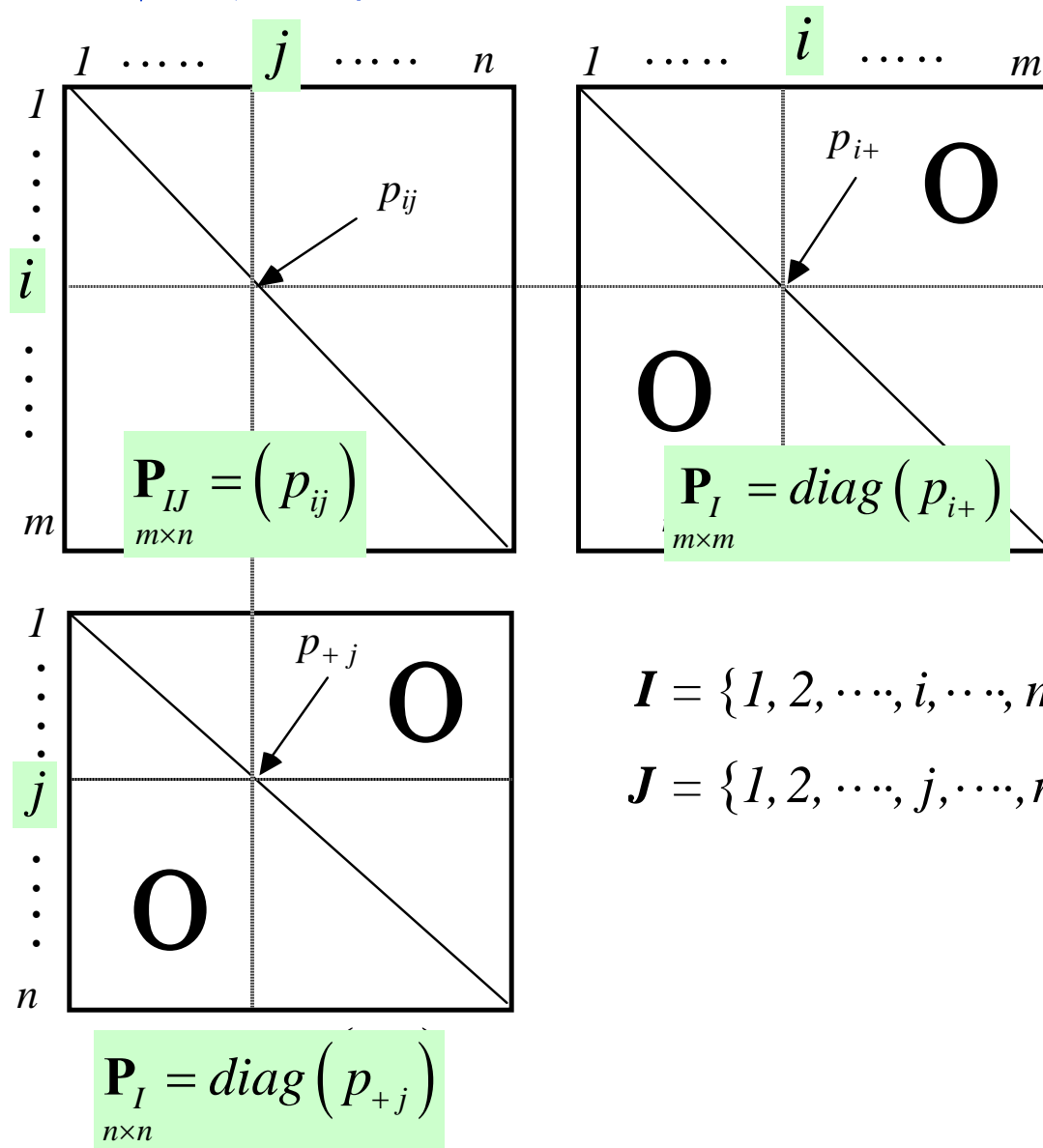
(上の各行列の要素)

★確認

$$\mathbf{P}_I = \underset{m \times m}{diag}(p_{i+}) = \begin{pmatrix} p_{1+} & & & \mathbf{O} \\ & p_{2+} & & \\ & & \ddots & \\ & & & p_{i+} & & \\ \mathbf{O} & & & & \ddots & \\ & & & & & p_{m+} \end{pmatrix} \quad \begin{array}{l} (m \times m \text{の対角行列}) \\ \text{行の質量が対角要素} \end{array}$$

$$\mathbf{P}_J = \underset{n \times n}{diag}(p_{+j}) = \begin{pmatrix} p_{+1} & & & \mathbf{O} \\ & p_{+2} & & \\ & & \ddots & \\ & & & p_{+j} & & \\ \mathbf{O} & & & & \ddots & \\ & & & & & p_{+n} \end{pmatrix} \quad \begin{array}{l} (n \times n \text{の対角行列}) \\ \text{列の質量が対角要素} \end{array}$$

図で表すと, ...



$$\mathbf{P}_{IJ} = (p_{ij})_{m \times n}$$

$$\mathbf{P}_I = \text{diag}(p_{i+})_{m \times m}$$

$$\mathbf{P}_J = \text{diag}(p_{+j})_{n \times n}$$

この3つの行列の関係
を覚えておこう.
イメージで覚える.

$$I = \{1, 2, \dots, i, \dots, m\}$$

$$J = \{1, 2, \dots, j, \dots, n\}$$

対応分析法の記法

- “プロフィール”を考えることが重要(再確認).
- 従来の統計学の言葉で言えば, “相対確率”のこと.
- これをより意味の深い, 拡張した見方で捉える.
- “行プロフィール”, “列プロフィール”がある.
- プロファイルの分布を“雲”(nuage, 英語のcloud)という.

$$\mathbf{N}_I = \mathbf{P}_I^{-1} \mathbf{P}_{IJ} = \left\{ q_{ij} = \frac{p_{ij}}{p_{i+}} \mid i \in I, j \in J \right\} \quad (\text{行のプロファイル})$$

$$\mathbf{N}_J = \mathbf{P}_J^{-1} \mathbf{P}_{JI} = \left\{ q_{ij}^* = \frac{p_{ij}}{p_{+j}} \mid i \in I, j \in J \right\} \quad (\text{列のプロファイル})$$

注: $\mathbf{P}_I^{-1}, \mathbf{P}_J^{-1}$ は, それぞれ $\mathbf{P}_I, \mathbf{P}_J$ の逆行列

\mathbf{P}_{JI} は, \mathbf{P}_{IJ} の転置行列 (\mathbf{P}_{IJ}^t)

プロフィールの特徴

- プロファイルとは行あるいは列の比率のパターンである.
- 一般に生データ(測定値, 度数) そのものではない.
- 対応分析では, 列や行の周辺和(質量)によって基準化する(長さをそろえる).
- この基準化操作は, 一般に, 周辺和が小さいものほど, 度数の変化に敏感に反応する.
- そのため, 列和や行和が小さい場合は, “はずれ値”などの影響を受けやすい.
- 調査データであると, 「無回答」「DK」, それと「どちらでもない」(ニュートラル)がはずれ値となりやすい.

(つづき)

- そうならないような“質問文”や“選択肢”の作り方が重要.
例1: 自記式調査(ウェブ, 郵送)では「どちらでも」はやめる
例2: 選択肢数をやたらに増やさない, 回答度数の散布を抑制.
- 自由回答データの分析などで, 2元データ表のセル内度数が“非常に少ない”, また“周辺和も度数が少ない”ような場合はかなりのはずれ値が生じる. 別の見方, 手当が必要.
- 初期の“2元データ表”の作り方, 集め方が重要.
- いろいろ手当は考えられてはいる(追加処理による, 一時的除去と再配置, subset CAの適用など).

再確認: ここで以下の関係に注意

- “プロフィール”とは2元クロス表の(多次元情報の)特徴を行あるいは列の比率のパターンで観察していること.
- 統計学の普通の言い方であると“相対確率”のこと.
- しかし, この形でデータを読むことに意味がある. 2元クロス表から, より一般的な“2元データ表”に拡張して利用.
- 下のプロフィールの関係を再確認する.

行のプロファイル

列のプロファイル

$$\begin{aligned} q_{ij} &= \frac{p_{ij}}{p_{i+}} = \frac{f_{ij}}{f_{i+}} \\ q_{ij}^* &= \frac{p_{ij}}{p_{+j}} = \frac{f_{ij}}{f_{+j}} \end{aligned}$$

★さらに以下を用意(形式的なこと)

p_{i+} 行の質量

$\mathbf{r}_{m \times 1}$ 列プロファイルの平均ベクトル(重心)

$$\mathbf{r}_{m \times 1} = (p_{1+}, p_{2+}, \dots, p_{i+}, \dots, p_{m+})^t \quad \text{または} \quad \mathbf{r}_{m \times 1} = \mathbf{P}_{m \times n} \mathbf{1}_n$$

p_{+j} 列の質量

$\mathbf{c}_{n \times 1}$ 行プロファイルの平均ベクトル(重心)

$$\mathbf{c}_{n \times 1} = (p_{+1}, p_{+2}, \dots, p_{+j}, \dots, p_{+n})^t \quad \text{または} \quad \mathbf{c}_{n \times 1} = \mathbf{P}_{n \times m}^t \mathbf{1}_m$$

クロス表, 確率行列を想起!

ここで, “質量”(mass), “重心”(center of gravity, centroid)は物理学の用語がある.

対応分析法では, こうした語句を好んで用いる(ベンゼクリの影響).

(つづき)

つぎの単位ベクトルを用意

$$\mathbf{1}_m = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{m \times 1}, \quad \mathbf{1}_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{n \times 1}$$

行の質量, 列の質量はこう表記できる

\mathbf{P}_{IJ} や \mathbf{P}_{JI} を使うとこう書ける

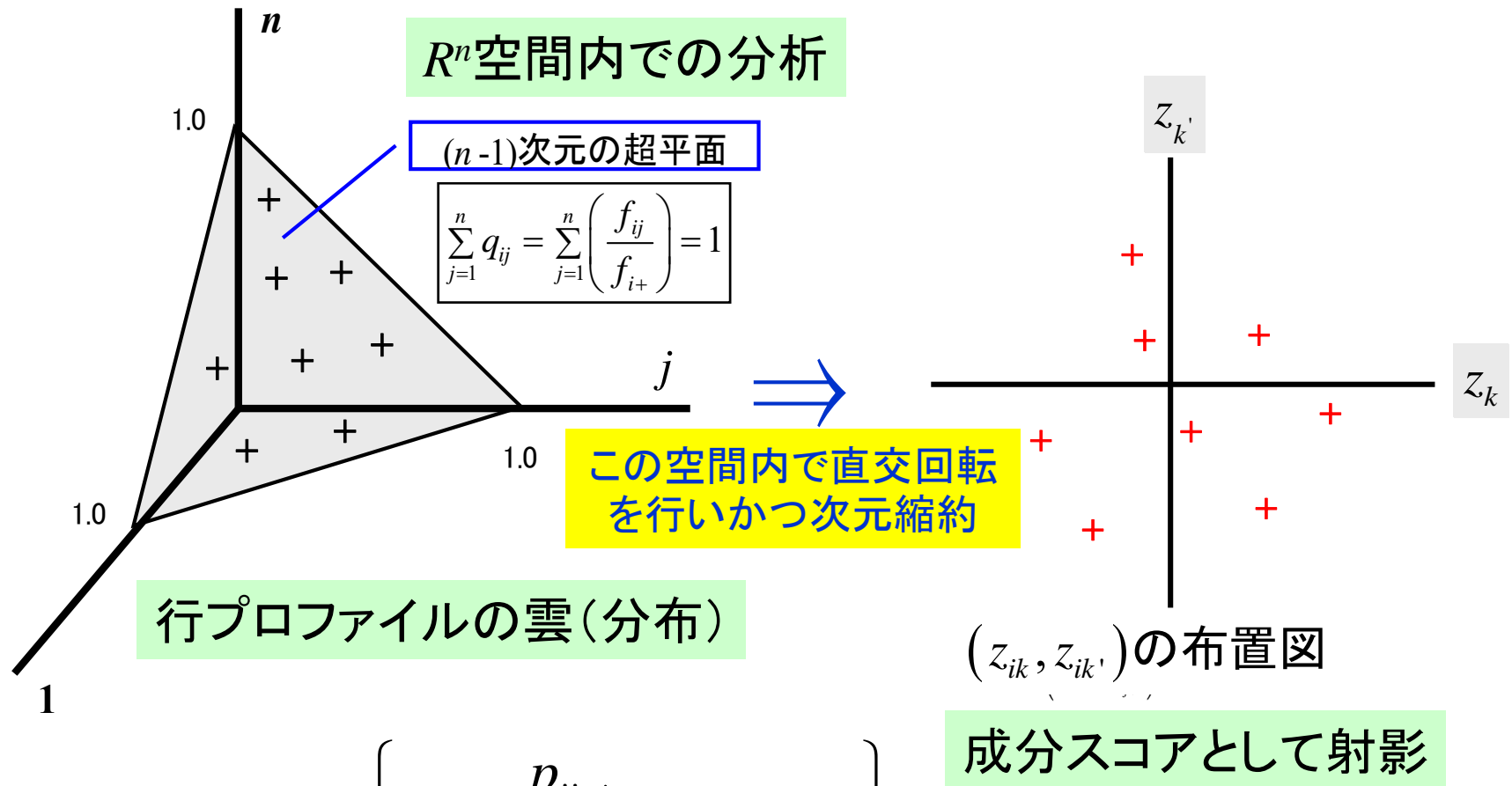
$$\mathbf{r} = \mathbf{P}_{IJ} \mathbf{1}_n = \begin{pmatrix} p_{1+} \\ p_{2+} \\ \vdots \\ p_{i+} \\ \vdots \\ p_{im} \end{pmatrix}_{m \times 1}, \quad \mathbf{c} = \mathbf{P}_{JI} \mathbf{1}_m = \begin{pmatrix} p_{+1} \\ p_{+2} \\ \vdots \\ p_{+j} \\ \vdots \\ p_{+n} \end{pmatrix}_{n \times 1}$$

行の質量

列の質量

ここらは, ベクトル, 行列を使うと, こう書ける
ことを知っておけばよい. のちの準備.

行プロファイルの分布(「行側」からの観察)

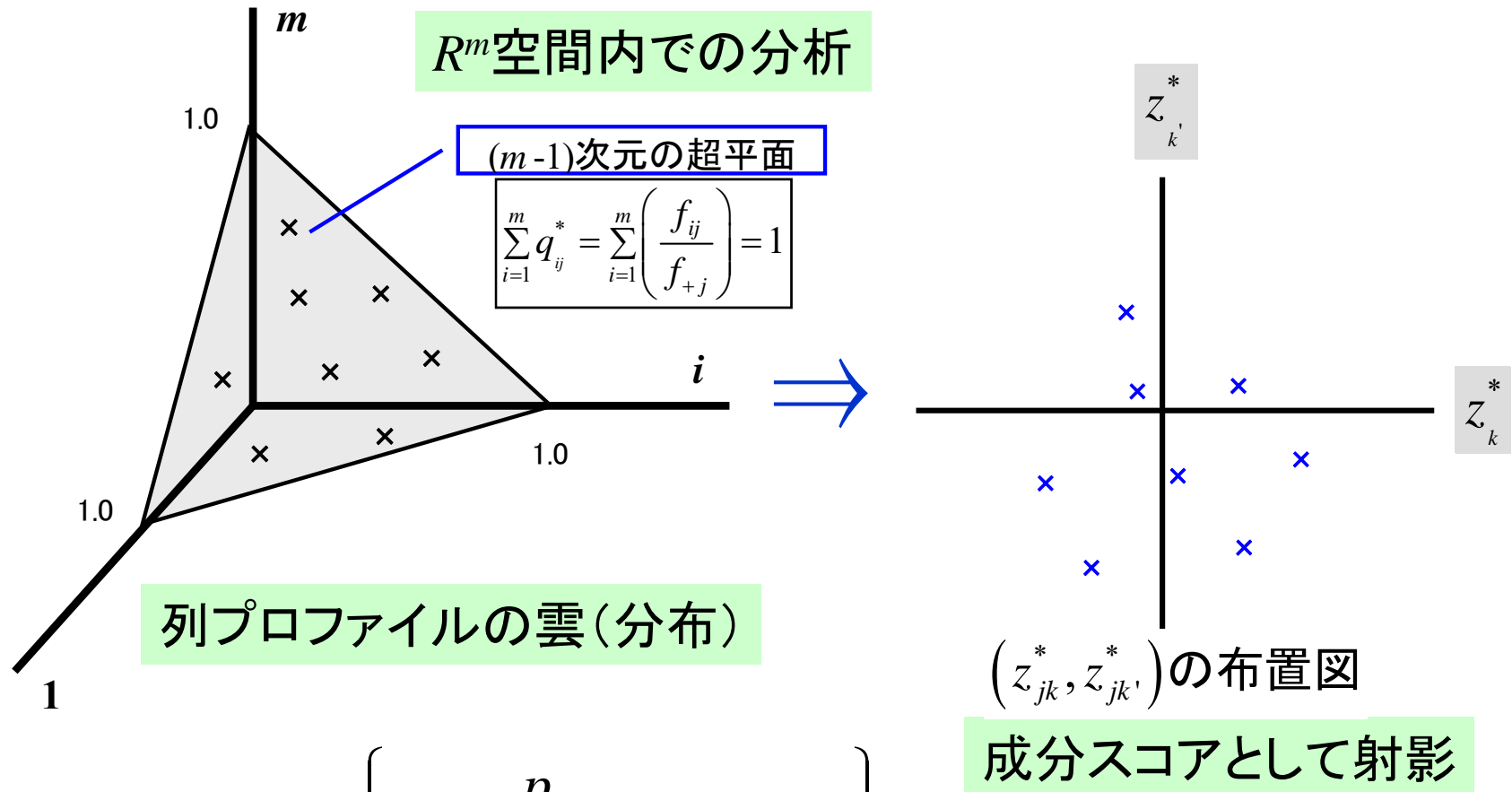


$$\mathbf{N}_I = \mathbf{P}_I^{-1} \mathbf{P}_{IJ} = \left\{ q_{ij} = \frac{p_{ij}}{p_{i+}} \mid i \in I, j \in J \right\}$$

$m \times n$ $m \times m$ $m \times n$

$(n-1)$ 次元空間に分布する m 個の点

列プロファイルの分布(「列側」からの観察)



$$\mathbf{N}_J = \mathbf{P}_J^{-1} \mathbf{P}_{J\mathbf{I}} = \left\{ q_{ij}^* = \frac{p_{ij}}{p_{+j}} \mid i \in I, j \in J \right\}$$

$n \times m$ $n \times n$ $n \times m$

$(m-1)$ 次元空間に分布する n 個の点

プロフィール空間内の縮約化

- 行あるいは列のプロファイルの分布は、多次元空間に布置するデータ(多次元データ)と考えられる.
- これを、より少数次元の空間内に“なるべく情報の損失なく射影”できるのか(次元縮約, 最適化の問題).
- これは、主成分分析などと同じ“**合成変数**”(合成指標化)に通底する考え方.
- 数理的には、“**固有値問題**”や“**特異値分解**”に帰着する方法である.
- 幾何学的には“(直交)回転”を行い、データの視点を変えること.
- ソフトが処理してくれること. ただしソフトが“何を出力したか”は知っておくべきこと.

(つづき)

- 対応分析法の特徴の1つは, 2元データ表の行と列とを対等にみていること,
- つまり“**対称**”に考えていること. (†)
- ここまでの説明では, 行側からみた, 列側(にある複数の変数)の合成変数を調べてきた.
- しかし, 行, 列のどちらから観察しても同等, つまり データ表を転置して解いても同じ解となること(そのように調整した).
- 別の例(トイ・データ)で, 何を行うかを確認する.

(†) 別の考え方もある(例: 非対称対応分析)

ここから別の例で確認しよう

- ある小さなトイ・データ(レストランの評価)を用いる.
- 2つの質問 I, J からなる調査で意見を聴取したと想定.
- 回答者数は, $N=1,284$ (人)であった.

質問I: 次にあげるレストランのうち, あなたがお気に入りのレストランはどれですか. (1つ選ぶ)

$$I = \{1, 2, 3, \dots, 10\}$$

- | | | | |
|---------|-----------|--------|---------|
| 1. いりふね | 2. かりや | 3. きくみ | 4. さとみ |
| 5. クラーク | 6. コルシカ | 7. バッハ | 8. ムガール |
| 9. ラ・マレ | 10. ロゴスキー | | |

質問J: そのレストランを選択したときの評価基準は, 次の3つのうちのどれでしょう. (1つ選ぶ)

$$J = \{1, 2, 3\}$$

- | | | |
|------------|------|------|
| 1. 工夫・サービス | 2. 味 | 3. 量 |
|------------|------|------|

回収した調査データのイメージ

項目 回答者	I (レストラン)	J (評価基準)
1	バッハ	味
2	ムガール	量
3	さとみ	量
4	ラ・マレ	工夫・サービス
5	きくみ	味
⋮	⋮	⋮
⋮	⋮	⋮
N	いりふね	量
$N=1,284$ (回答者数)		

なんども繰り返すが、これは“質的データ”であり、選択肢という“標識”を示している。

2元クロス表: $\mathbf{F}_{10 \times 3} = (f_{ij})$

$m = 10, n = 3 \Rightarrow I = \{1, 2, \dots, 10\}, J = \{1, 2, \dots, 3\}$

項目I \ 項目J	工夫・サービス	味	量	行和
いりふね	98	25	32	155
かりや	105	35	38	178
きくみ	35	8	67	110
さとみ	42	46	7	95
クラーク	34	14	54	102
コルシカ	32	77	13	122
バッハ	48	76	18	142
ムガール	49	44	16	109
ラ・マレ	49	82	15	146
ロゴスキー	48	35	42	125
列和	540	442	302	1,284 (=N)

行プロファイルの算出例

$$\mathbf{N}_I = \mathbf{P}_I^{-1} \mathbf{P}_{IJ} = \left\{ q_{ij} = \frac{p_{ij}}{p_{i+}} \mid i \in I, j \in J \right\}$$

これをすべてのセルについて算出.
列プロファイルも同じように求められる.

項目I \ 項目J	工夫 サービス	味	量	行和
いりふね	98/155=0.632	85/155=0.161	32/155=0.206	1
かりや	105/178=0.590	35/178=0.197	38/178=0.213	1
きくみ	35/110=0.318	8/110=0.073	67/110=0.609	1
さとみ	42/95=0.442	1
クラーク	1
コルシカ	1
バツハ	1
ムガール	1
ラ・マレ	1
ロゴスキー	48/125=0.384	35/125=0.280	42/125=0.336	1
列和	540/1284 =0.421	442/1284 =0.344	302/1284 =0.235	1

行プロフィール: $\mathbf{N}_I = \mathbf{P}_I^{-1} \mathbf{P}_{IJ} = \left\{ q_{ij} = \frac{p_{ij}}{p_{i+}} \mid i \in I, j \in J \right\}$

評価項目 レストラン	工夫 サービス	味	量
いりふね	0.632	0.161	0.206
かりや	0.590	0.197	0.213
きくみ	0.318	0.073	0.609
さとみ	0.442	0.484	0.074
クラーク	0.333	0.137	0.529
コルシカ	0.262	0.631	0.107
バッハ	0.338	0.535	0.127
ムガール	0.450	0.404	0.147
ラ・マレ	0.336	0.562	0.103
ロゴスキー	0.384	0.280	0.336
p_{+j} (質量)	0.421	0.344	0.235
行プロフィールの重心 (列の質量)	$\mathbf{c} = (0.421, 0.344, 0.235)^t$		

行和はすべて「1」

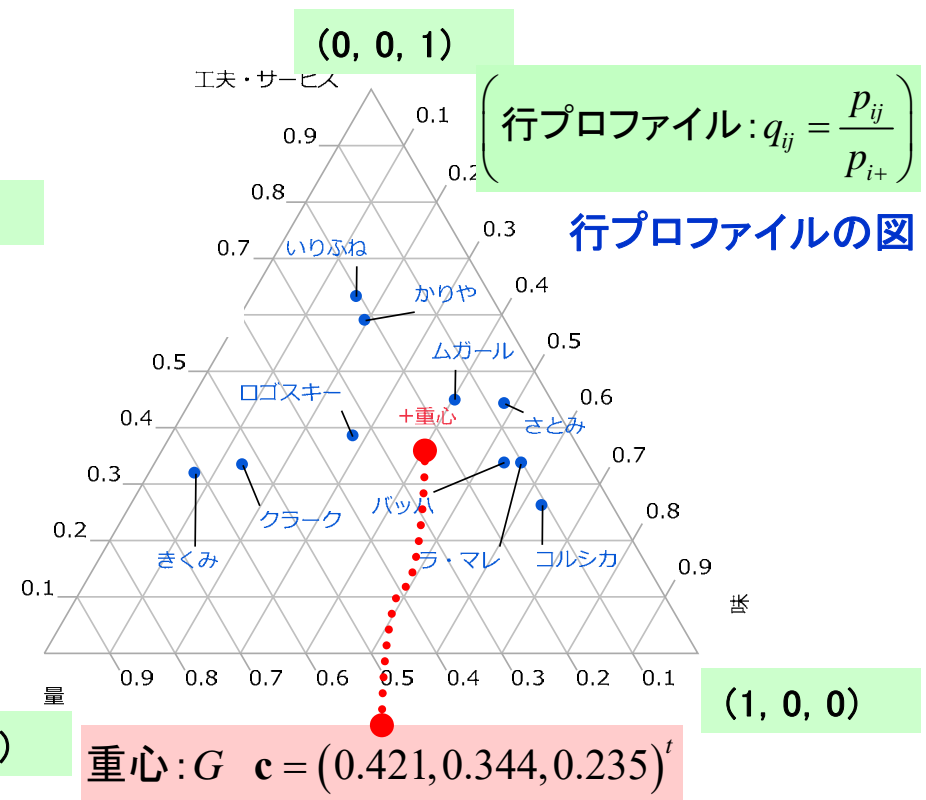
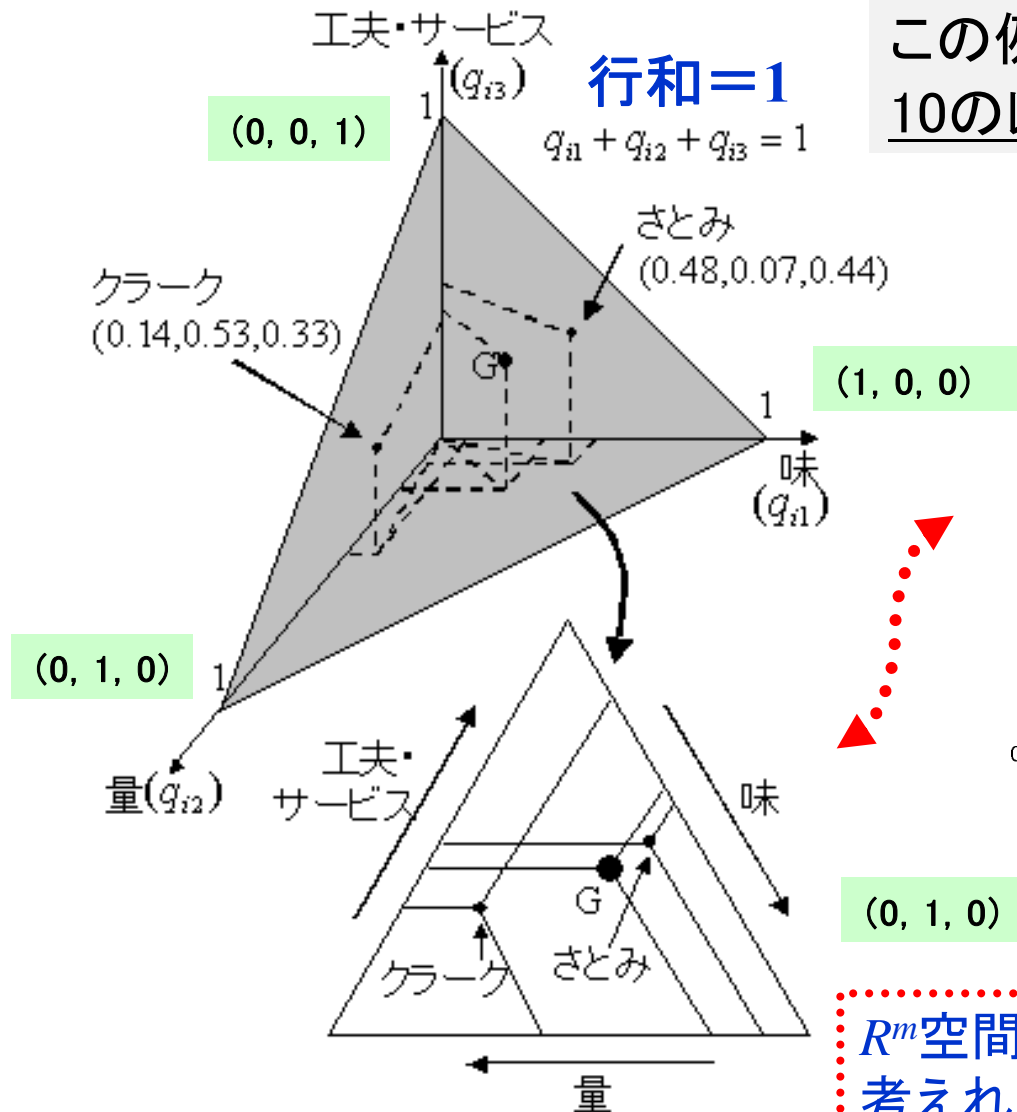
列プロフィール: $\mathbf{N}_J = \mathbf{P}_J^{-1} \mathbf{P}_{JI} = \left\{ q_{ij}^* = \frac{p_{ij}}{p_{+j}} \mid i \in I, j \in J \right\}$

評価項目 レストラン	工夫 サービス	味	量	p_{i+} (質量)	列プロフィールの重心 (行の質量)
いりふね	0.181	0.057	0.106	0.121	$\mathbf{r} = \begin{pmatrix} 0.121 \\ 0.139 \\ 0.086 \\ 0.074 \\ 0.079 \\ 0.095 \\ 0.111 \\ 0.085 \\ 0.114 \\ 0.097 \end{pmatrix}$
かりや	0.194	0.079	0.126	0.139	
きくみ	0.065	0.018	0.222	0.086	
さとみ	0.078	0.104	0.023	0.074	
クラーク	0.063	0.032	0.179	0.079	
コルシカ	0.059	0.174	0.043	0.095	
バツハ	0.089	0.172	0.06	0.111	
ムガール	0.091	0.1	0.053	0.085	
ラ・マレ	0.091	0.186	0.05	0.114	
ロゴスキー	0.089	0.079	0.139	0.097	

列和はすべて「1」

R^n 空間内の分析(行の側から)を考える

この例では $(n-1)=3-2=2$ 次元(平面)
10のレストランが3つの評価基準に分布



R^m 空間内での分析も同じように
考えれば良い

“行の側”から観察

- クロス表の[行和] = 1ということから, 3つの「評価項目」のうち, 自由に動かせるのは2つの列まで. 下の式.
- この例では(自由度が1つ減って)2次元の平面内で行プロファイルは“分布”している.
- この分布を“雲”(nuage, 英語のcloud)と名付けた.
- この2次元平面はアミカケ部となり, これを映した“三角図”(三角座標系)の各辺の長さは「1」である(比率だから).

$$q_{i1} + q_{i2} + q_{i3} = 1 \quad (i = 1, 2, \dots, 10) \quad (\text{アミをかけた平面})$$

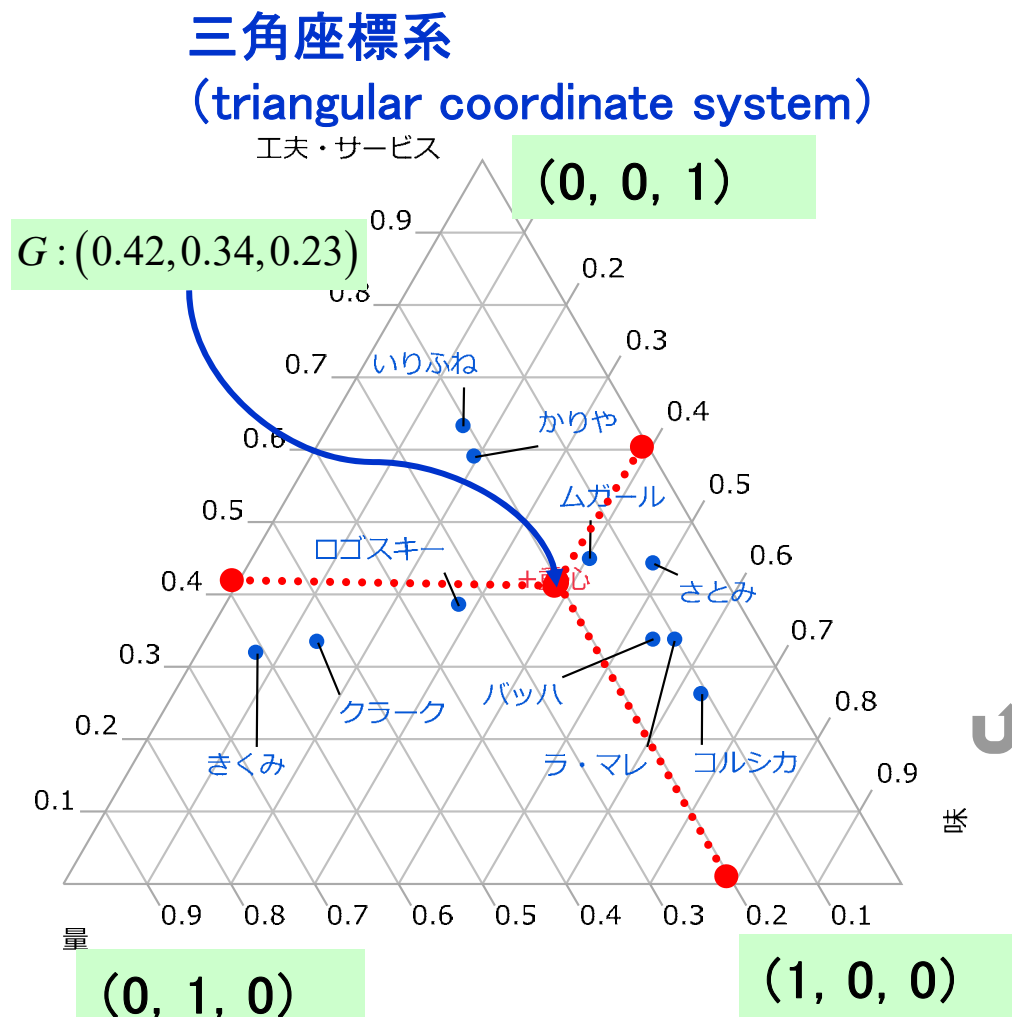
$$G = (0.421, 0.344, 0.235) \quad (\text{重心, つまり行平均ベクトル})$$

$$\text{これに同じ: } \mathbf{c} = (0.421, 0.344, 0.235)^t \quad (\text{列の質量})$$

三角図(三角座標系)による表示と特徴

- ここでは, 全情報が平面, つまり**三角座標系** (triangular coordinate system) で表示される.
- 一般には, 2元データ表の寸法は大きくなるので, 概念的には $(n-1)$ 次元の超平面を考えことになる.
- そのような座標系を“**重心座標系**” (barycentric coordinate system) という.
- これをいま, データ要素の単位 (ストレッチ・プロフィール) で描いてみる.

$$x_{ij} = \frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} = \frac{q_{ij}}{\sqrt{p_{+j}}}$$



これを図で描くとどうなるのか？

ストレッチ・プロファイルの意味

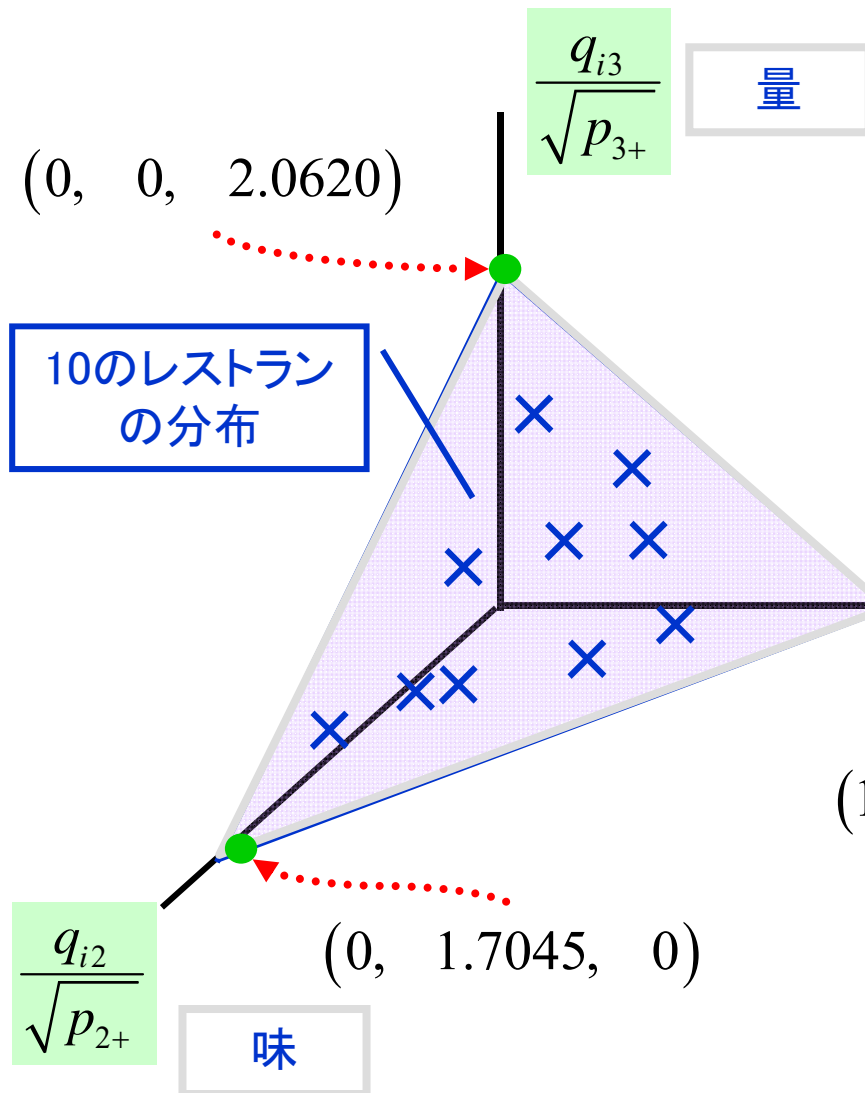
- まず, ストレッチの重みを算出する.

$$\begin{aligned}
 \mathbf{c}_{n \times 1} = \begin{pmatrix} \sqrt{p_{+1}} \\ \sqrt{p_{+2}} \\ \vdots \\ \sqrt{p_{+j}} \\ \vdots \\ \sqrt{p_{+n}} \end{pmatrix} &\Rightarrow \mathbf{c}_{n \times 1}^{-1} = \begin{pmatrix} \frac{1}{\sqrt{p_{+1}}} \\ \frac{1}{\sqrt{p_{+2}}} \\ \vdots \\ \frac{1}{\sqrt{p_{+j}}} \\ \vdots \\ \frac{1}{\sqrt{p_{+n}}} \end{pmatrix} \\
 &\Rightarrow \mathbf{c}_{3 \times 1} = \begin{pmatrix} \sqrt{p_{+1}} \\ \sqrt{p_{+2}} \\ \sqrt{p_{+3}} \end{pmatrix} \Rightarrow \mathbf{c}_{3 \times 1}^{-1} = \begin{pmatrix} \frac{1}{\sqrt{p_{+1}}} \\ \frac{1}{\sqrt{p_{+2}}} \\ \frac{1}{\sqrt{p_{+3}}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{0.4206}} \\ \frac{1}{\sqrt{0.3442}} \\ \frac{1}{\sqrt{0.2352}} \end{pmatrix} \\
 &= \begin{pmatrix} 1.5419 \\ 1.7045 \\ 2.0620 \end{pmatrix} \begin{matrix} \leftarrow \text{工夫・サービス} \\ \leftarrow \text{味} \\ \leftarrow \text{量} \end{matrix}
 \end{aligned}$$

一般には, ...
(n 個の選択肢)

レストランデータでは, ...
(3個の「評価基準」選択肢)

各辺を伸ばす



正三角形系 \Rightarrow ストレッチ系

$$(1, 0, 0) \Rightarrow (1.5419, 0, 0)$$

$$(0, 1, 0) \Rightarrow (0, 1.7045, 0)$$

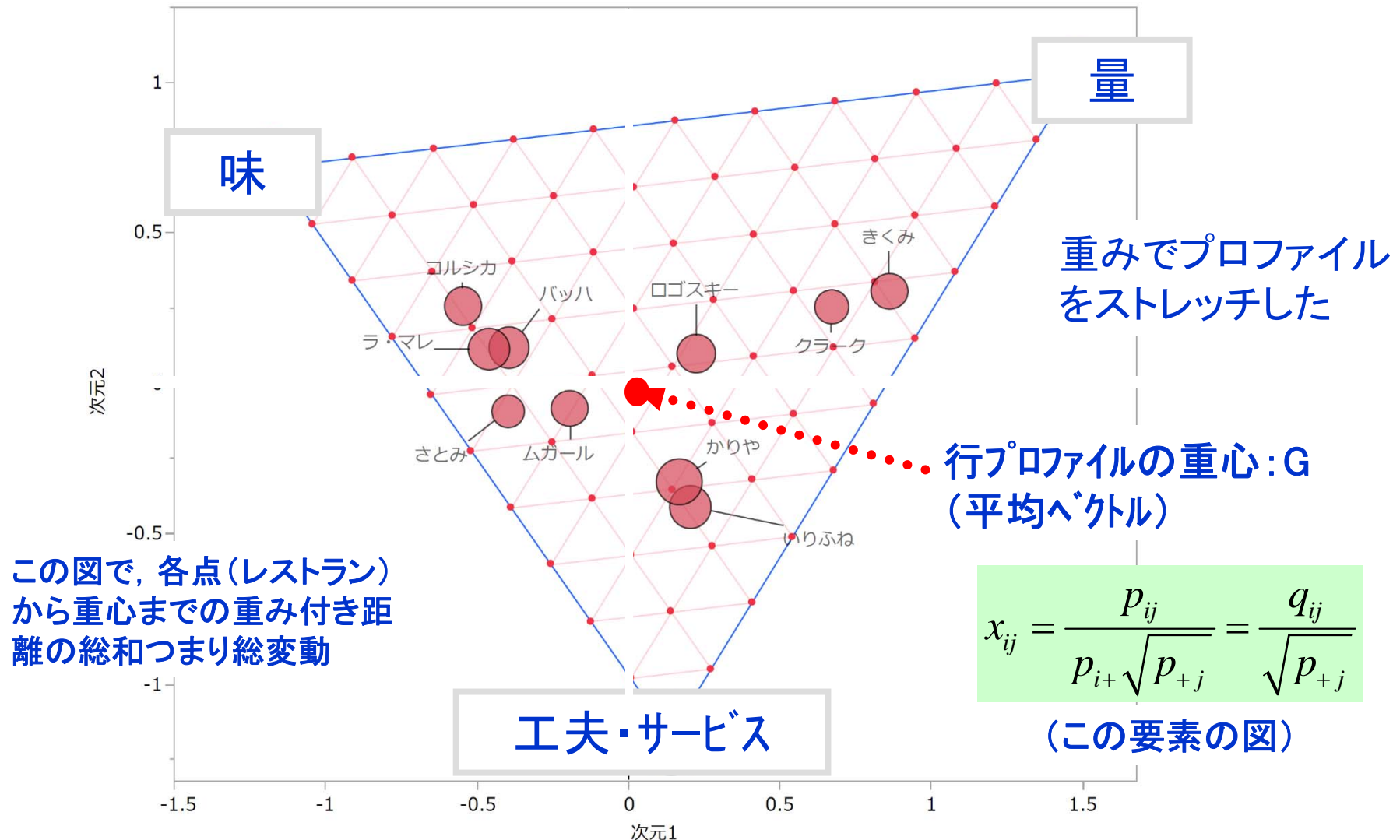
$$(0, 0, 1) \Rightarrow (0, 0, 2.0620)$$

この分布!!!
(これがデータ)

$$x_{ij} = \frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} = \frac{q_{ij}}{\sqrt{p_{+j}}}$$

成分スコアの式を思い出す!!!

ストレッチ・プロファイルの図



ここでカイ二乗距離の役割を調べる

- 対応分析の特徴の1つは“**カイ二乗距離**”を用いること.
- プロファイルあるいはストレッチ・プロファイルを考えることと, 密接に関係する.
- プロファイルの分布の様子(変動, チラバリ)を測る指標として“**(平方)カイ二乗距離**”を用いる.
- 距離は変動を測る指標であること. (平方)距離を考えることは, 変動(分散)を考えることである.
- 総変動(全慣性)との関係があること.
- 分布の“**同等性**”を保持すること.

(つづき)

- ユークリッド距離を用いると, 2元データ表(クロス表)の行和あるいは列和の大きさの違いが反映されないこと.
- カイ二乗距離とユークリッド距離の関係(プロフィール間のカイ二乗距離は成分スコアのユークリッド距離).
- この性質は“クラスター化処理”で重要であること.
- プロファイル間のカイ二乗距離やプロフィールと重心との距離を, このカイ二乗距離で評価する.
- 利点はいろいろある. そのいくつかを調べる.

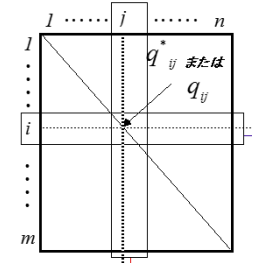
“カイ二乗距離”の用意(1)

- プロファイル間の距離を測る指標として“(平方)カイ二乗距離”を考えること. 距離は変動の指標でもある.
- ある重み(加重)付の(平方)ユークリッド距離である.
- ではその“重み”はなにか, これが重要な意味をもつ.
- その重みとして“質量”を使ってみる.
- 得られる(平方)カイ二乗距離は単純な“(平方)ユークリッド距離”とどう違うのか. (つまり重みの役割は)

カイ二乗距離(chi-square distance), 平方カイ二乗距離ということもある. 質量(mass), ユークリッド距離・平方ユークリッド距離.

カイ二乗距離の用意(2)

- ここで, (平方)カイ二乗距離を $d_B^2(\cdot, \cdot)$ で表す. ベンゼクリの提案になる距離という意味を含めて添字「 B 」を使う.
- 行プロファイル間, 列プロファイル間のカイ二乗距離を以下のように書く.



行プロファイル i と i' 間の“(平方)カイ二乗距離”

$$d_B^2(i, i') = \sum_{j=1}^n \frac{1}{p_{+j}} (q_{ij} - q_{i'j})^2 = \sum_{j=1}^n \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2$$

(加重が “ $\frac{1}{p_{+j}}$ ”)

列プロファイル j と j' 間の“(平方)カイ二乗距離”

$$d_B^2(j, j') = \sum_{i=1}^m \frac{1}{p_{i+}} (q_{ij}^* - q_{ij'}^*)^2 = \sum_{i=1}^m \frac{1}{p_{i+}} \left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}} \right)^2$$

(加重が “ $\frac{1}{p_{i+}}$ ”)

(質量の逆数)

確認

- つぎのようにも書けることに注意する. ストレッチ・プロファイルはどこにあるか？

行プロファイル*i*と*i'*間の“(平方)カイニ乗距離”

$$d_B^2(i, i') = \sum_{j=1}^n \frac{1}{p_{+j}} (q_{ij} - q_{i'j})^2 = \sum_{j=1}^n \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2$$

加重が“ $\frac{1}{p_{+j}}$ ”

$$= \sum_{j=1}^n \left(\frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} - \frac{p_{i'j}}{p_{i'+} \sqrt{p_{+j}}} \right)^2 = \sum_{j=1}^n \left(\frac{q_{ij}}{\sqrt{p_{+j}}} - \frac{q_{i'j}}{\sqrt{p_{+j}}} \right)^2$$

列プロファイル*j*と*j'*間の“(平方)カイニ乗距離”

重み付きプロファイル

$$d_B^2(j, j') = \sum_{i=1}^m \frac{1}{p_{i+}} (q_{ij}^* - q_{ij'}^*)^2 = \sum_{i=1}^m \frac{1}{p_{i+}} \left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}} \right)^2$$

加重が“ $\frac{1}{p_{i+}}$ ”

$$= \sum_{i=1}^m \left(\frac{p_{ij}}{p_{+j} \sqrt{p_{i+}}} - \frac{p_{ij'}}{p_{+j'} \sqrt{p_{i+}}} \right)^2 = \sum_{i=1}^m \left(\frac{q_{ij}^*}{\sqrt{p_{i+}}} - \frac{q_{ij'}^*}{\sqrt{p_{i+}}} \right)^2$$

(つづき)

- 正の平方根を作れば, “カイ二乗距離”となる.
- これを $d_B^2(\cdot, \cdot)$ と書くことにする. (添字「B」を付ける)
- 重みがなければ, 単にプロファイル間の“ユークリッド距離”となる.
- 相対度数を使った, 以下の右の式の関係にも注意する.
- ある種の確率分布の距離になっている.

$$d_B(i, i') = \sqrt{\sum_{j=1}^n \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2} = \sqrt{\sum_{j=1}^n \frac{1}{p_{+j}} \left(\frac{f_{ij}}{f_{i+}} - \frac{f_{i'j}}{f_{i'+}} \right)^2}$$

$$d_B(j, j') = \sqrt{\sum_{i=1}^m \frac{1}{p_{i+}} \left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}} \right)^2} = \sqrt{\sum_{i=1}^m \frac{1}{p_{i+}} \left(\frac{f_{ij}}{f_{+j}} - \frac{f_{ij'}}{f_{+j'}} \right)^2}$$

カイ二乗距離とユークリッド距離の違い

- プロファイル間の平方ユークリッド距離を $d_E^2(\cdot, \cdot)$ と書くと、これらは以下となる。（添字「E」を付ける）
- この正の平方根がユークリッド距離となる。
- 常に“カイ二乗距離 > ユークリッド距離”となること。
- プロファイル間のカイ二乗距離は、成分スコア間のユークリッド距離となる。重要な性質（後述）

行プロファイル i と i' 間の“平方ユークリッド距離”

$$d_E^2(i, i') = \sum_{j=1}^n (q_{ij} - q_{i'j})^2 = \sum_{j=1}^n \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2 = \sum_{j=1}^n \left(\frac{f_{ij}}{f_{i+}} - \frac{f_{i'j}}{f_{i'+}} \right)^2$$

列プロファイル j と j' 間の“平方ユークリッド距離”

$$d_E^2(j, j') = \sum_{i=1}^m (q_{ij}^* - q_{ij'}^*)^2 = \sum_{i=1}^m \left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}} \right)^2 = \sum_{i=1}^m \left(\frac{f_{ij}}{f_{+j}} - \frac{f_{ij'}}{f_{+j'}} \right)^2$$

常に“カイ二乗距離＞ユークリッド距離”

- ここはプロフィール間の距離の比較のこと.
- これは以下からわかる.
- では, なぜこのようなカイ二乗距離を使うのか?
- ここに重要な意味がある. これをまた数値例で追ってみる.

① p_{+j}, p_{i+} は確率であるから, $0 \leq p_{+j}, p_{i+} \leq 1$ となる.

② その値で割ることから, つねに $\frac{1}{p_{+j}}, \frac{1}{p_{i+}} \geq 1$

③ よって, つねに “ $d_B^2(\cdot, \cdot) > d_E^2(\cdot, \cdot)$ ” である.

ここまでに見たことは“何を行っていた”のか?

注意: 後述のように, “ストレッチ・プロフィール”間の(平方)カイ二乗距離は, “成分スコア”間の(平方)ユークリッド距離に等しい.

要約1: “行プロフィール”側からの観察

行の側から分析

n 次元空間, R^n 内での分析を行う

行和を1としたときの「行のプロファイル」を $(n-1)$ 次元内に分布する m 個の点と考える

行プロフィールの分布(行の“雲”)の分析

$$\mathbf{N}_I = \mathbf{P}_I^{-1} \mathbf{P}_{IJ} = \left\{ q_{ij} = \frac{p_{ij}}{p_{i+}} \mid i \in I, j \in J \right\}$$

行のプロファイル i と i' 間の“(平方)カイ二乗距離”

$$d_B^2(i, i') = \sum_{j=1}^n \frac{1}{p_{+j}} (q_{ij} - q_{i'j})^2 = \sum_{j=1}^n \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2$$

要約2: “列プロフィール”側からの観察

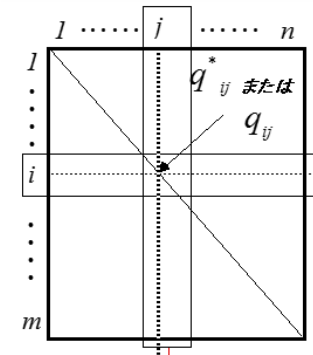
列の側から分析
m 次元空間, R^m 内での分析を行う
列和を1としたときの「列のプロファイル」を $(m-1)$ 次元内に分布する n 個の点と考える
<p>列プロフィールの分布(行の“雲”)の分析</p> $\mathbf{N}_J = \mathbf{P}_J^{-1} \mathbf{P}_{JI} = \left\{ q_{ij}^* = \frac{p_{ij}}{p_{+j}} \mid i \in I, j \in J \right\}$
<p>列のプロファイル j と j' 間の“(平方)カイ二乗距離”</p> $d_B^2(j, j') = \sum_{i=1}^m \frac{1}{p_{i+}} (q_{ij}^* - q_{ij'}^*)^2 = \sum_{i=1}^m \frac{1}{p_{i+}} \left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}} \right)^2$

行側(レストラン)の距離

- 行側(レストラン)に注目して, レストラン間のユークリッド距離とカイ二乗距離を比べる.
- いくつかの組み合わせについて調べよう.
- 「さとみ」と「きくみ」「ムガール」「いりふね」の距離.

$$d_E^2(i, i') = \sum_{j=1}^n (q_{ij} - q_{i'j})^2 = \sum_{j=1}^n \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2 \quad (\text{平方)ユークリッド距離}$$

$$d_B^2(i, i') = \sum_{j=1}^n \frac{1}{p_{+j}} (q_{ij} - q_{i'j})^2 = \sum_{j=1}^n \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2 \quad (\text{平方)カイ二乗距離}$$



$$d_E^2(\text{さとみ}, \text{きくみ}) = \sum_{j=1}^3 \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2$$

$$= \left(\frac{46}{95} - \frac{8}{110} \right)^2 + \left(\frac{7}{95} - \frac{67}{110} \right)^2 + \left(\frac{42}{95} - \frac{35}{110} \right)^2 = 0.4705$$

単純なプロフィール間の
ユークリッド距離

$$d_E^2(\text{さとみ}, \text{ムガール}) = 0.0118, d_E^2(\text{さとみ}, \text{いりふね}) = 0.1579$$

$$d_B^2(\text{さとみ}, \text{きくみ}) = \sum_{j=1}^3 \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2$$

こちらはカイ二乗距離

$$= \frac{1}{442/1284} \left(\frac{46}{95} - \frac{8}{110} \right)^2 + \frac{1}{302/1284} \left(\frac{7}{95} - \frac{67}{110} \right)^2 + \frac{1}{540/1284} \left(\frac{42}{95} - \frac{35}{110} \right)^2$$

$$= 1.7456$$

$$d_B^2(\text{さとみ}, \text{ムガール}) = 0.0414, d_B^2(\text{さとみ}, \text{いりふね}) = 0.4632$$

さとみ	42/95	46/95	7/95
きくみ	35/110	8/110	67/110
ムガール	49/109	44/109	16/109
いりふね	98/155	25/155	32/155
質量	540/1284	442/1284	302/1284

留意点は, ...

- 列和の逆数, つまり“質量の逆数”を重みとした加重ユークリッド距離つまり“カイ二乗距離”としたことの効果はなにか.
- 列和の大小が距離に反映される.
- 列和が小さいと, 度数の少しの差が大きく評価される.
- 列和が大きいと, 度数の差が大きくても距離への影響が少ない(抑える).
- 行と列を入れ換えても, つまり「列和」→「行和」と読み替えても同じ性質.

(つづき)

- カイ二乗距離とすることで, ほかの性質も生まれる.
- **分布の同等性**: 行あるいは列のパターンの同等性があること. (うしろで例をみる)
- **距離の関係**: もとのクロス表から得たカイ二乗距離が, 成分スコアのユークリッド距離となること(非常に重要).
- 数値例として求めた2種の距離の結果は以下.

組み合わせ	平方ユークリッド距離	カイ二乗距離
「さとみ」と「きくみ」	0.4705	1.7456
「さとみ」と「ムガール」	0.0118	0.0414
「さとみ」と「いりふね」	0.1579	0.4632

列側(評価)についても同じように求められる

$$d_E^2(j, j') = \sum_{i=1}^m (q_{ij}^* - q_{ij'}^*)^2 = \sum_{i=1}^m \left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}} \right)^2$$

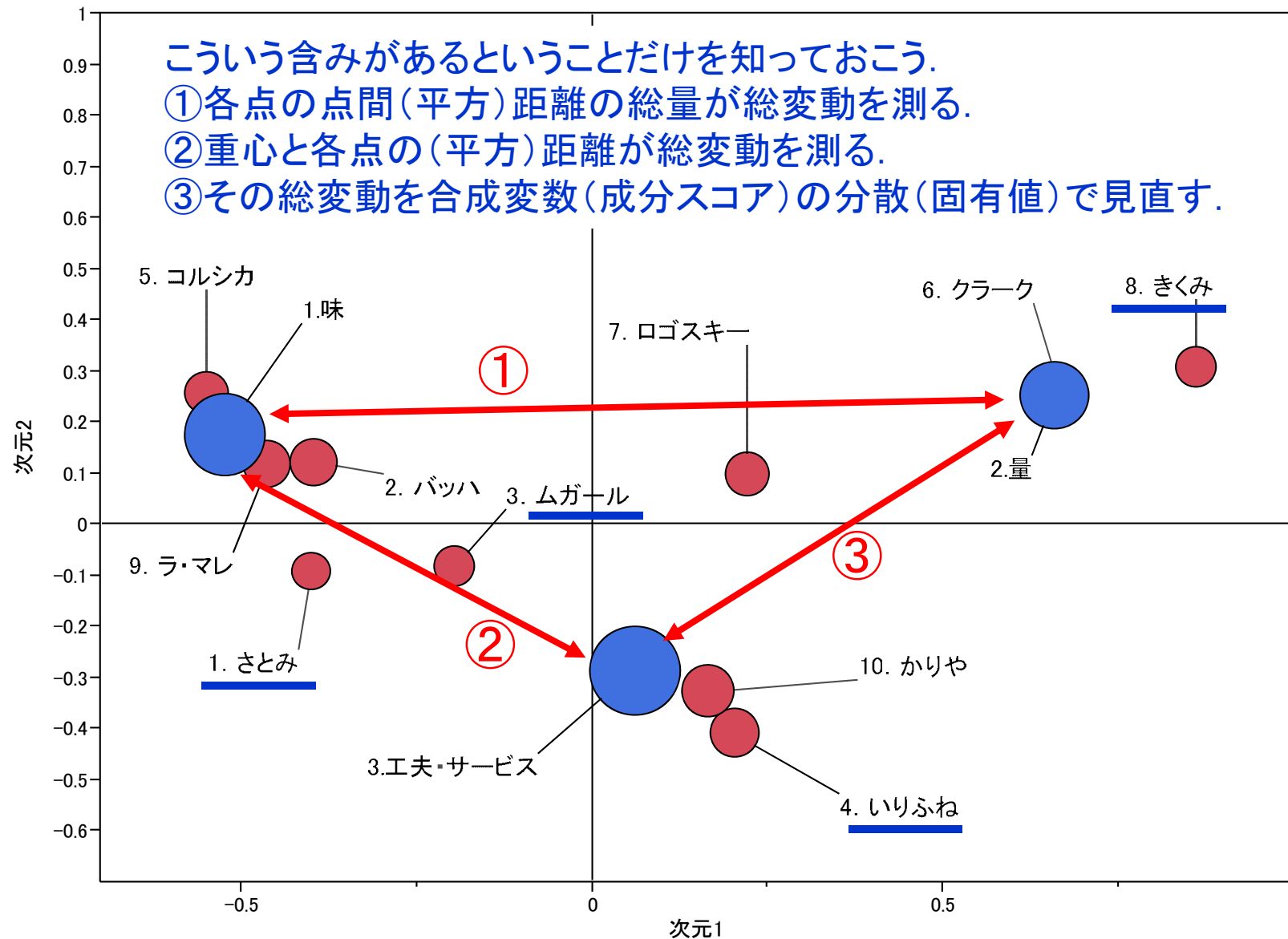
プロフィール間の
平方ユークリッド距離

$$d_B^2(j, j') = \sum_{i=1}^m \frac{1}{p_{i+}} (q_{ij}^* - q_{ij'}^*)^2 = \sum_{i=1}^m \frac{1}{p_{i+}} \left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}} \right)^2$$

プロフィール間の
平方カイ二乗距離

	平方ユークリッド距離		(平方)カイ二乗距離	
	量	工夫・サービス	量	工夫・サービス
味	① 0.128406	② 0.061767	① 1.401066	② 0.553017
量	0	③ 0.058101	0	③ 0.645842

縦横の尺度を揃えた同時布置図で確認



総変動(全慣性)と成分スコアの分散(固有値)

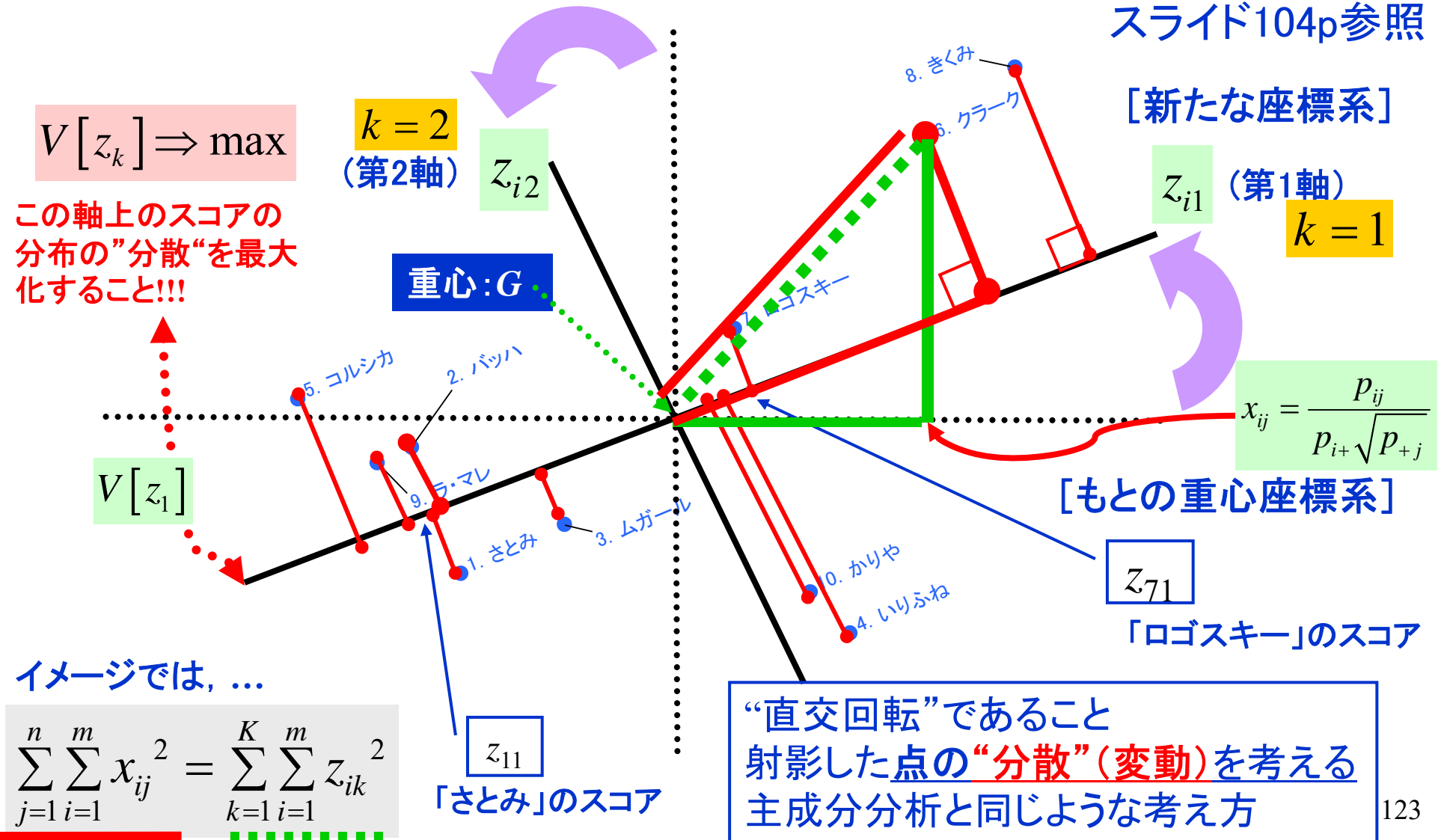
- ここで、総変動(全慣性)と成分スコアの分散(固有値)の関係を模式的に(再)確認する.
- 目標は、2元データ表全体の総変動を、どのように分解するかである.
- はじめのデータの空間、つまりプロフィール空間(雲)、つまり“重心座標系”の中での変動を観察.
- これを別の空間、つまり成分スコアの作る新たな座標系へと射影する.
- そしてなるべくなら、元の空間次元よりも少ない次元内に納めたい.

(つづき)

- いわゆる次元の縮約(“次元縮約”)を行うことになる.
- 同時に“無相関化”すること(直交回転を行う).
- 数理的には, 空間の回転操作, つまり“行列の分解”を行うこと. (重心の周りで回転)
- この点で, 主成分分析, 次元縮約のある判別分析, (一部の)クラスター化法などに類似する.
- 要点は“ストレッチ・プロファイルの空間”(重心座標系)の点の布置図空間を新たな成分スコアの空間(合成変数系)に移す(射影)すること.

重心座標系から (z_1, z_2) 系へ変換(射影)

スライド104p参照



ここで何を行うか？[ことばで書いてみる]

- むずかしく考えない，図の上で以下のように考える.
 - ①ある回転で射影した(仮の)“スコア”(z)の分散を作る.
 - ②まず，初めの軸(第1軸)のスコアの分散を最大化するような向きを探す. これを“第1主軸”(first principal axis)という. これが記号で「 $k=1$ 」に対応.
 - ③つぎに分散の大きい第2軸の向きを，第1軸に直交するように決める. これを“第2主軸”(second principal axis)とする. つまり「 $k=2$ 」に対応.
 - ④この例では，2つの軸までしかない. 一般にはさらに“高次元”であるから，順に主軸を探す. つまり， $k=1,2,3,\dots$ と探す. 分散が大きい順，つまり固有値の大きい順.

(つづき)

- 数理的にみると, “(直交)回転を行う”こと. これは“総変動の分解”であり点(選択肢, ここはレストラン)の分布の“視点を変える”ことに相当する. ということと容認して....
- “固有値問題”を解くことに同じこと. これが通常の方法.
- 特異値分解としても同様に考えられる(前述の数値例).
- 重要なことは, 元の2元データ表が保有する“総変動(全慣性)”を見方を変えて固有値の大きさで切り分けたこと.

(つづき)

- 元の固有値の大きさ＝成分スコアの分散，総変動（全慣性）＝固有値の総和であった（何度も例でみた）．
- つまり重心座標系の点の分布の総変動（全慣性）に注目し，これを別の合成変数の作る空間（成分スコア）に射影し切り分ける．
- 成分スコアの分散の大きさの順に重要度（寄与の程度，情報量）が低減する，ということ（そう考える）．
- そうなるように空間内の変動の大きさの向きを探した，ということ（回転の操作）．

重要な性質(全慣性の幾何学的解釈)

- プロファイルから重心までのカイ二乗距離, 質量, 全慣性の間には, 重要な関係がある.

総変動(全慣性)と固有値, カイ二乗統計量の関係(再確認)

$$(\text{全慣性}) = (\text{固有値の総和}) = \sum_{k=1}^K \lambda_k = \frac{\chi_p^2}{N} \quad [\text{2元データ表が保有する総変動}]$$

行プロファイルについて

$$\text{全慣性} = \sum_{i=1}^m \left[\underline{(\text{第}i\text{番目の質量})} \times (\text{第}i\text{番プロファイルから重心までのカイ二乗距離}) \right]$$

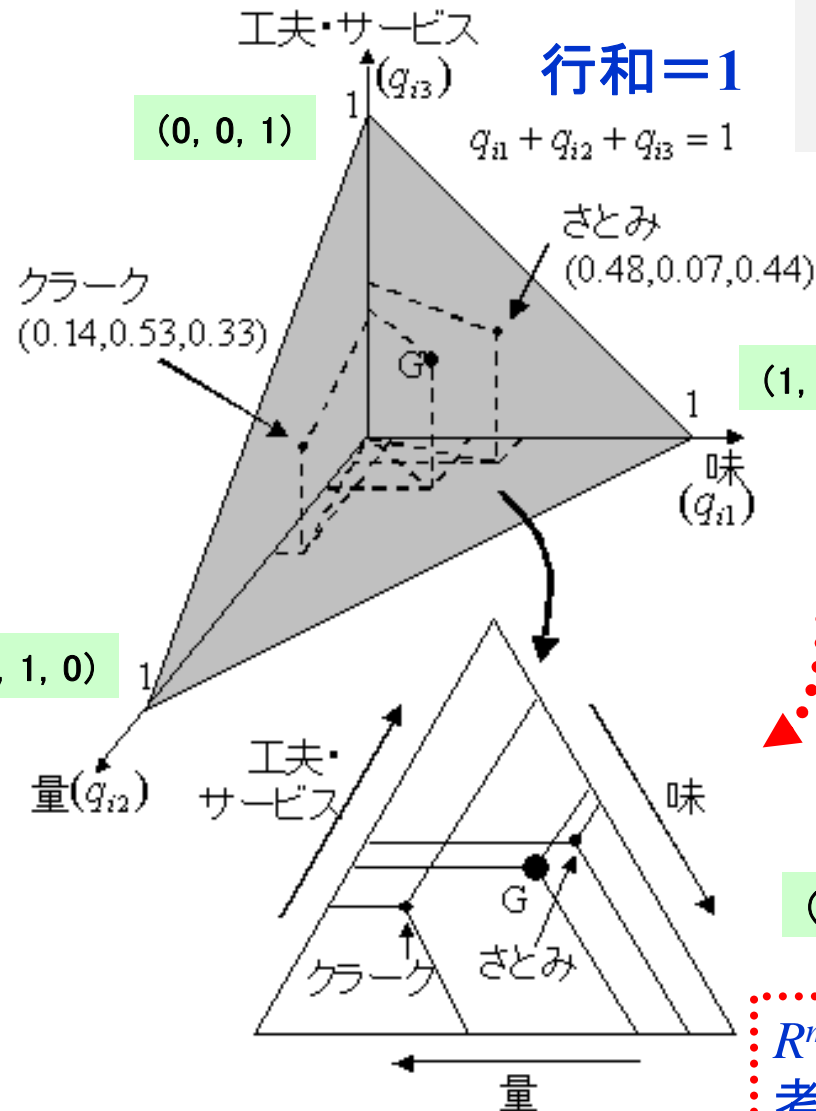
列プロファイルについて



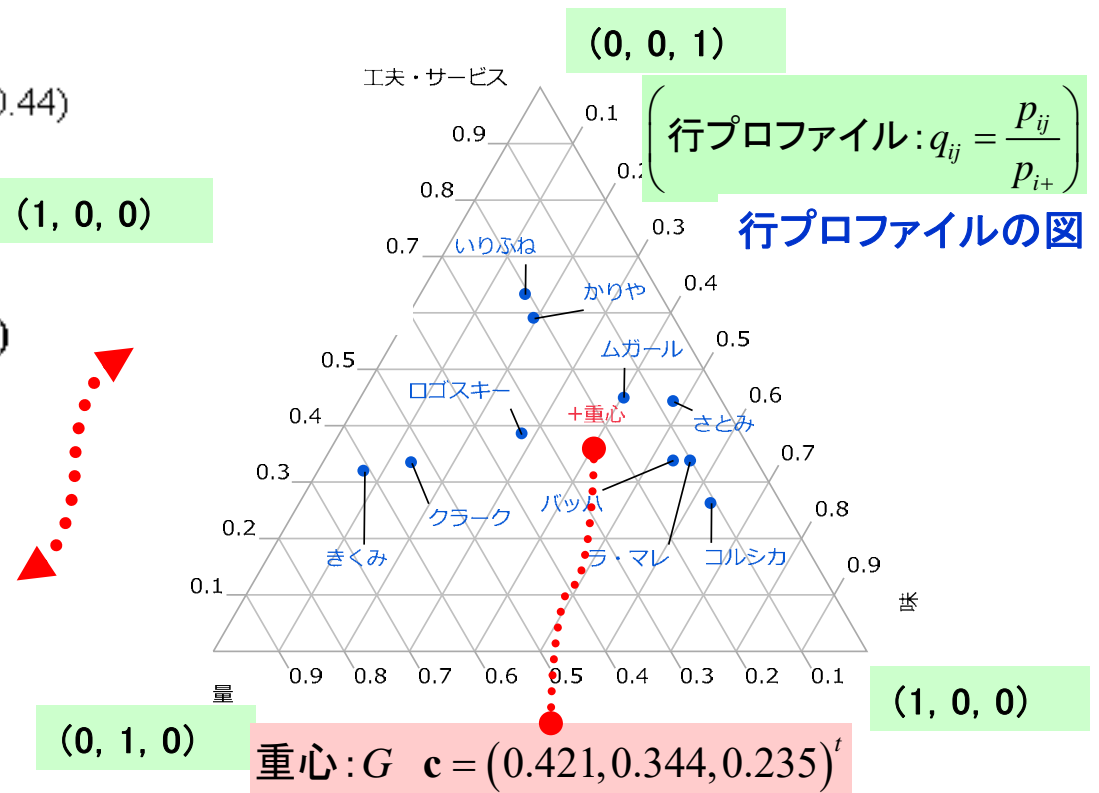
$$\text{全慣性} = \sum_{j=1}^n \left[\underline{(\text{第}j\text{番目の質量})} \times (\text{第}j\text{番プロファイルから重心までのカイ二乗距離}) \right]$$

通常 of データ解析で 分散 を求めることを想起する
どこか違うか, 質量の加重 がある. うしろの三角座標で確認.

再確認: R^n 空間内での分析(行側から)



この例では $(n-1)=3-2=2$ 次元(平面)
10のレストランが3つの評価基準に分布



R^m 空間内での分析も同じように
考えれば良い

再確認 & あらためて“行の側”から観察

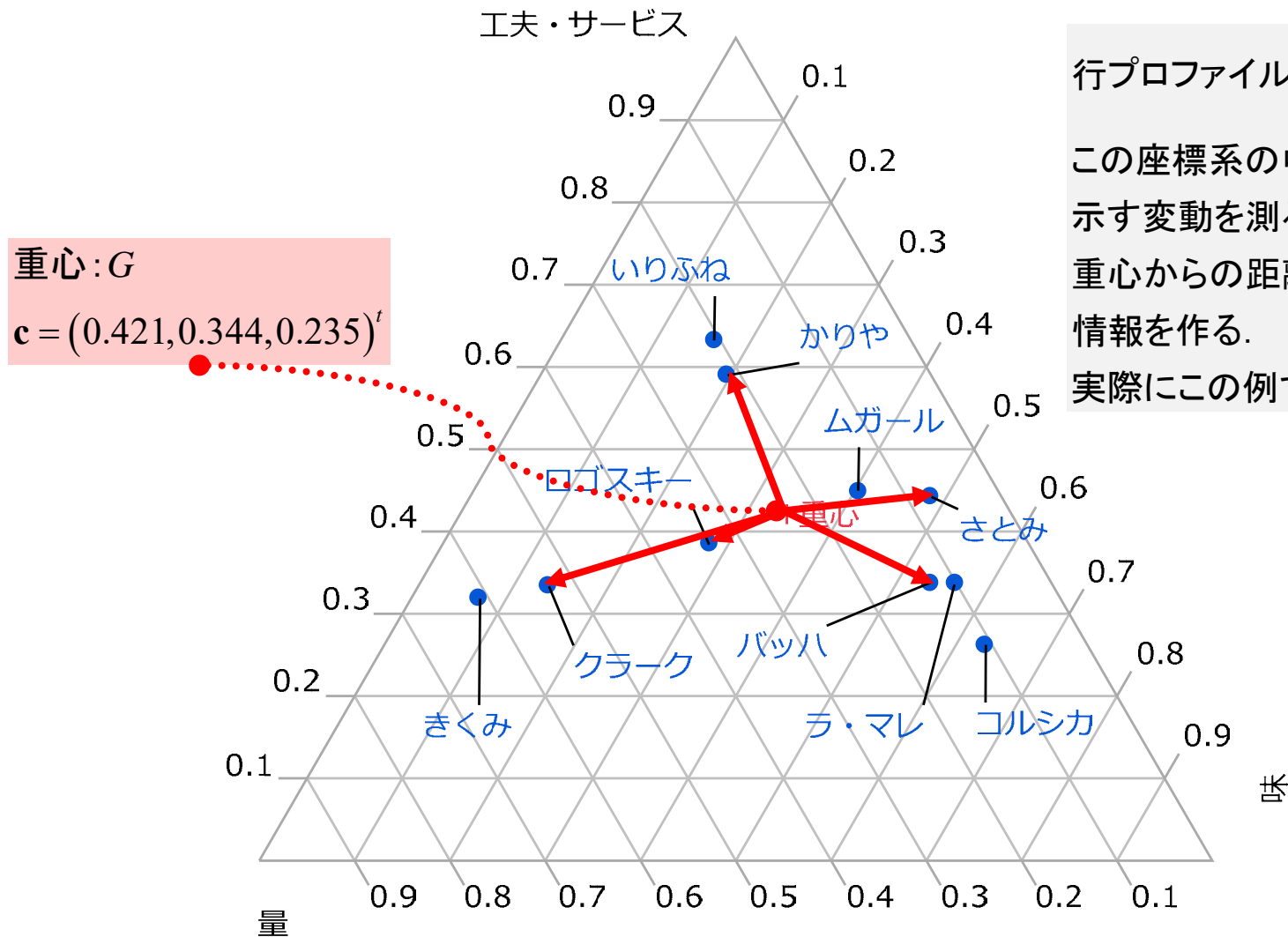
- クロス表の[行和] = 1ということから, 3つの「評価項目」のうち, 自由に動かせるのは2つの列まで. 下の式.
- この例では(自由度が1つ減って)2次元の平面内で行プロフィールは“分布”(雲)している. 普通はもっと高次元となる.
- 重心から各点(行の要素, レストラン)への平方カイ二乗距離を調べる.
- 次ページの図で確認する. これは三角座標系だが実際はストレッチ・プロフィール空間.

$$q_{i1} + q_{i2} + q_{i3} = 1 \quad (i = 1, 2, \dots, 10) \quad (\text{アミをかけた平面} = 2\text{次元})$$

$$G = (0.421, 0.344, 0.235) \quad (\text{重心, つまり行プロフィールの平均})$$

$$\mathbf{c} = (0.421, 0.344, 0.235)^t \quad (\text{つまり列の質量のこと})$$

「行プロファイル」のプロットで観察



行プロファイル: $q_{ij} = \frac{p_{ij}}{p_{i+}}$ をプロット

この座標系の中で各点(レストラン)が
示す変動を測る.

重心からの距離, つまり分散相当の
情報を作る.

実際にこの例でそれを作る(算出)する.



観察

- 点(レストラン)が, 重心から遠く離れる程, 距離が遠い, つまり変動(チラバリ)が大きい. 総変動が大きくなる.
- 重心からの“(平方)カイ二乗距離”を測ることは“分散”を調べることに同じ.
- また, 点がどういう方向, 向きに散布しているかが点(レストラン)の相対的な関係(どの向きで近い, 遠い)を示している.
- この空間内の“**総変動(全慣性)**”が, 固有値の総和となる(プロフィールを使う理由). 情報の総量と読む.
 - 例: 点が重心から広く散らばれば全慣性は大きくなる.
 - 例: 点が重心の周りに近く分布すれば, 全慣性は小さい.

このデータ表の固有値ほかの再確認

- このクロス表の対応分析法で得た固有値他を示す.
- ここでいま, 考えている空間は“2次元(平面)”である.
- これが $K=\min\{m,n\}-1=3-1=2$ の意味.

$$\left(\begin{array}{c} \text{全慣性} \\ \text{総変動} \end{array} \right) = (\text{固有値の総和}) = \sum_{k=1}^K \lambda_k = \frac{\chi_p^2}{N}$$

$$\sum_{k=1}^K \lambda_k = \lambda_1 + \lambda_2 = 0.19766 + 0.06001 = 0.25768 \quad (\blacktriangle)$$

$$v_1 = 76.7(\%), \quad v_2 = 23.3(\%) \quad (\text{寄与率})$$

数値例:「行側」から全慣性を求めてみる

- たとえば「さとみ」の行プロフィールについて調べる.
- 以下, 各レストラン(項目*l*)について重み付きの重心との間のカイ二乗距離を求め和を作る. 図で確認(赤い矢印).

$$\left[\begin{array}{l} \text{(`さとみ'の質量)} \\ \times (\text{'さとみ'プロフィールから行重心までのカイ二乗距離}) \end{array} \right]$$

$$= 0.074 \times \left\{ \underbrace{\frac{(0.442 - 0.421)^2}{0.421}}_{\text{エ夫・サービス}} + \underbrace{\frac{(0.484 - 0.344)^2}{0.344}}_{\text{味}} + \underbrace{\frac{(0.074 - 0.235)^2}{0.235}}_{\text{量}} \right\} = 0.01246$$

$$\sum_{i=1}^{10} \left[\begin{array}{l} \text{(第}i\text{番目の質量)} \\ \times (\text{第}i\text{番プロフィールから重心までのカイ二乗距離}) \end{array} \right]$$

$$\underbrace{0.02501}_{\text{いりふね}} + \underbrace{0.01818}_{\text{かりや}} + \underbrace{0.07172}_{\text{きくみ}} + \cdots + \underbrace{0.02616}_{\text{ラ・マレ}} + \underbrace{0.00568}_{\text{ログスキー}} = 0.25759 \quad (\blacktriangle)$$

項目*l*の各選択肢(レストラン)の変動の程度⇒合わせると全慣性



数値例:「列側」から全慣性を求めてみる

- 同様に列の選択肢「エ夫・サービス」について調べる.
- 以下, 各評価基準(項目 J)について, 重み付きの重心との間のカイ二乗距離を求め和を作る.

$$\left[\begin{array}{l} \text{(`エ夫・サービス'の質量)} \\ \times (\text{'エ夫・サービス'プロフィールから列重心までのカイ二乗距離}) \end{array} \right]$$

$$= 0.421 \times \left\{ \begin{array}{l} \frac{(0.181 - 0.121)^2}{0.121} + \frac{(0.194 - 0.137)^2}{0.114} + \dots \\ + \frac{(0.091 - 0.114)^2}{0.114} + \frac{(0.089 - 0.097)^2}{0.097} \end{array} \right\} = 0.03611$$

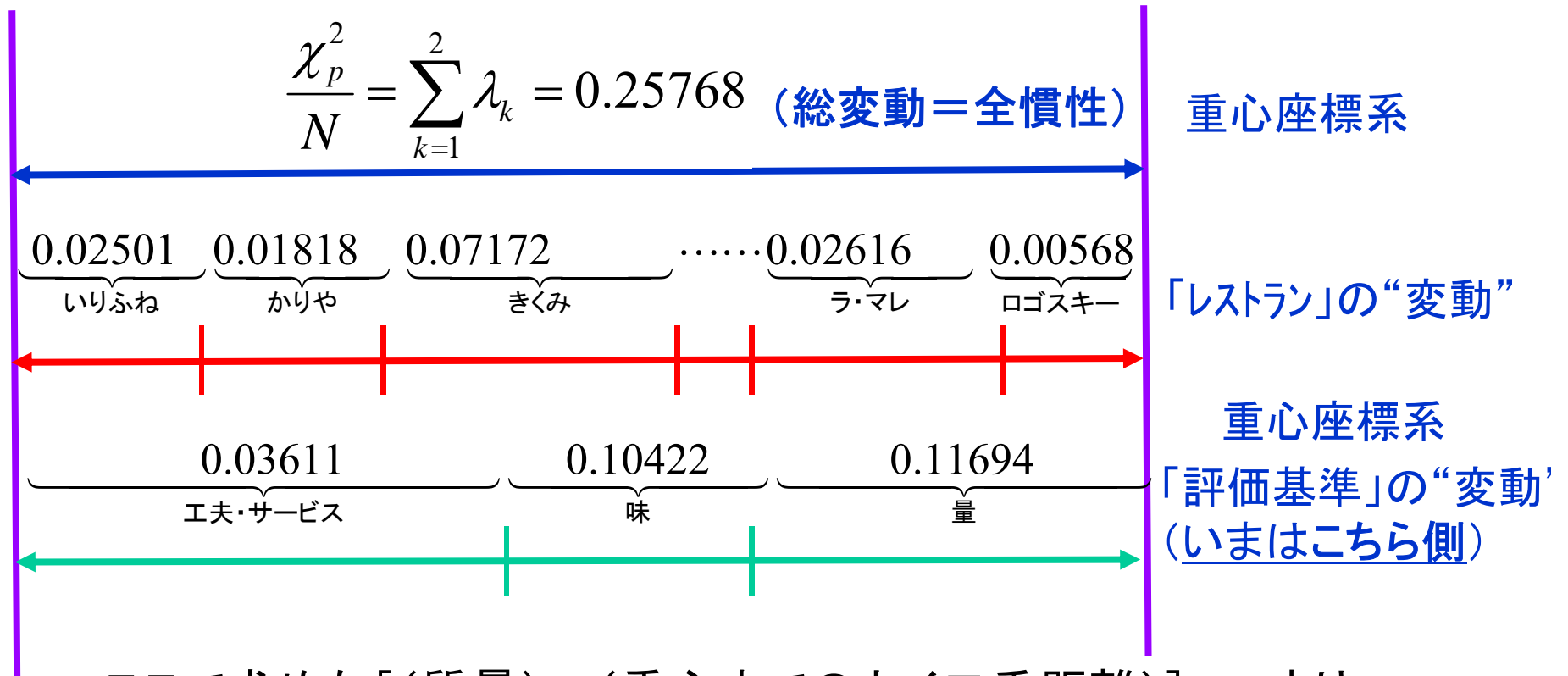
いりふね かりや
ラ・マレ ログスキー

$$\sum_{j=1}^3 \left[(\text{第}j\text{番目の質量}) \times (\text{第}j\text{番プロフィールから重心までのカイ二乗距離}) \right]$$

$$= \underbrace{0.03611}_{\text{エ夫・サービス}} + \underbrace{0.10422}_{\text{味}} + \underbrace{0.11694}_{\text{量}} = 0.25727 \quad (\blacktriangle)$$

項目 J の各選択肢(評価基準)の変動の程度
⇒ 合わせると全慣性

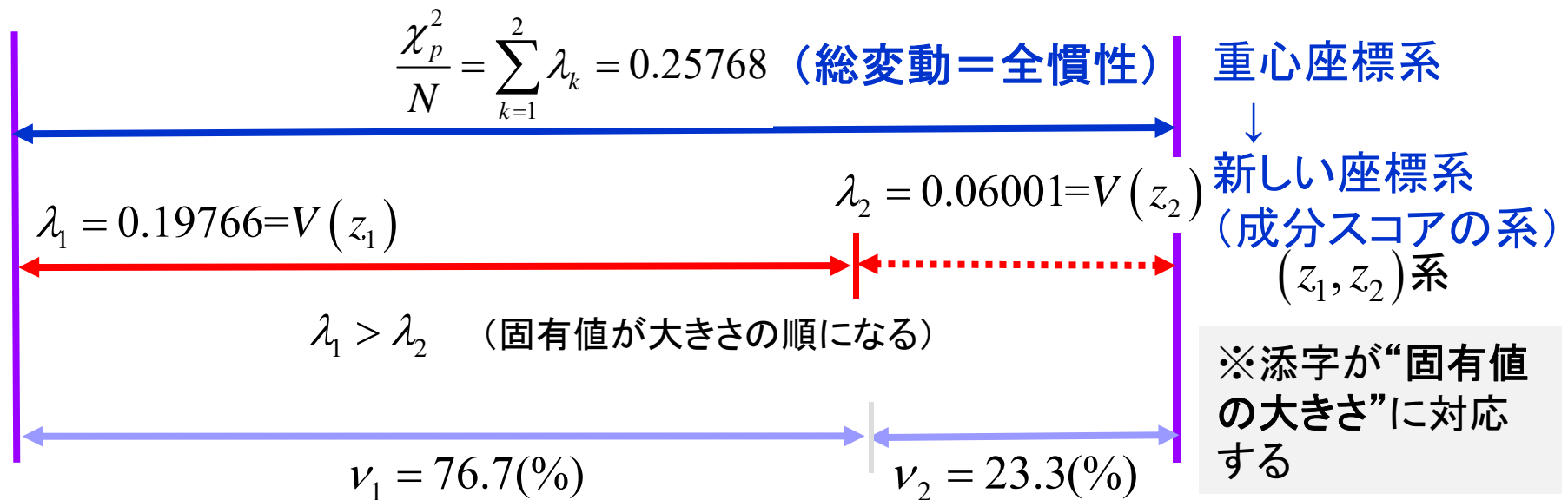
総変動(全慣性)と行・列の変動の関係



- ここで求めた[(質量) × (重心までのカイニ乗距離)], つまり個々の選択肢の変動を引用し図に描いた.
- 行(レストラン)と列(評価基準)の両側にある.
- 全体の和は”全慣性“に一致する(総変動という情報を分解).

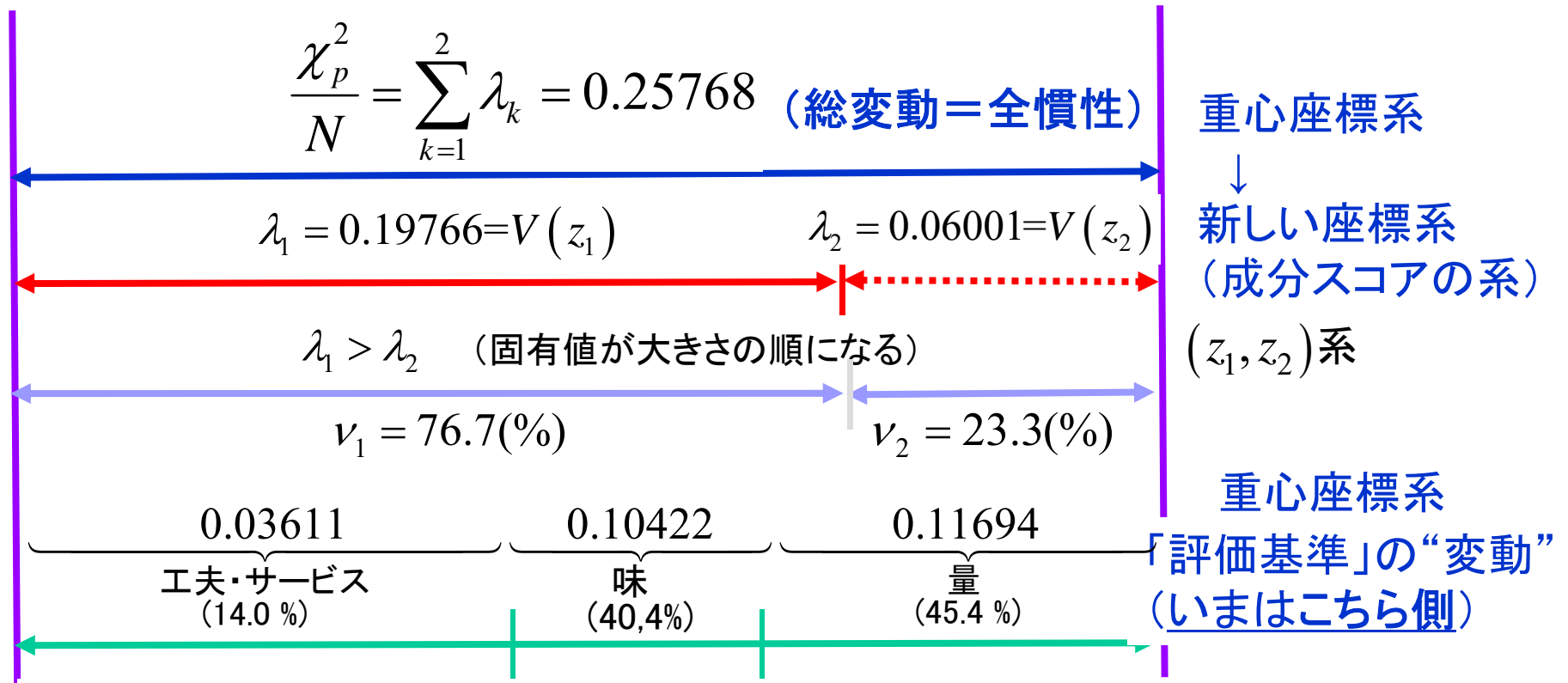


総変動(全慣性)と2つの固有値(分散)の関係



全体を固有値(成分スコアの分散かつ慣性)がどのように分割するか⇒寄与の大きさ(寄与率)として測る

総変動と列側（評価基準）の変動の関係



- ここでは、行側（レストラン）の点からみた列側（評価基準）の空間内の分布と考えている.
- 全体の和は”全慣性“に一致する（総変動という情報を分解）.

意味解釈は, ...

- 2つ(第1成分, 第2成分)の行成分スコアは“合成変数”である. 前に述べた式を思い出す.
- ここで(行成分スコアは)“3つの(列の)「評価基準」の加重和”である.
- 個々の成分スコアの中に, この3つの「評価基準」の情報が(合成変数として)含まれていることに注意する.
- 3つの評価項目の合わせた変動の総和をあらたな成分スコアの分散(固有値)がどう分けるかをみた.
- 固有値(分散, 慣性)の大きさの順に, 成分スコアの説明力は低減する.

(つづき)

- 総変動(全慣性, 固有値の総和)を全情報量とすると, あらたに生成した合成変数の第1固有値(第1分散)で全情報の76.7(%), 約80%を占める.
- つまりもっとも大きい第1成分スコアの変動でほとんど10のレストランの識別は可能となる.
- 残り(23.3%)を犠牲にして良ければ, 1次元(1成分)に縮約. つまり, 1次元で近似できる. 次元が縮約した.
- 通常は2元データ表の寸法(m と n)はかなり大きいから, 複数の成分(合成変数)を観察する.

この例で“第1成分スコア”について再確認

$$z_{ik} = \sum_{j=1}^n l_{jk} x_{ij} = \sum_{j=1}^n \left(\frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} \right) l_{jk} = \sum_{j=1}^n \left(\frac{q_{ij}}{\sqrt{p_{+j}}} \right) l_{jk} \quad (i \in I; k = 1, 2, \dots, K)$$

ここで、 $k=1$ とし、行側(10のレストラン)のスコアを示す。
行選択肢 $i \in I$ の第1成分スコアという合成変数は以下。

$$z_{i1} = \sum_{j=1}^3 l_{j1} x_{ij} = \underline{l_{11}} x_{i1} + \underline{l_{21}} x_{i2} + \underline{l_{31}} x_{i3}$$

$$= \underbrace{\frac{p_{i1}}{p_{i+} \sqrt{p_{+1}}}}_{\text{工夫・サービス}} \underline{l_{11}} + \underbrace{\frac{p_{i2}}{p_{i+} \sqrt{p_{+2}}}}_{\text{味}} \underline{l_{21}} + \underbrace{\frac{p_{i3}}{p_{i+} \sqrt{p_{+3}}}}_{\text{量}} \underline{l_{31}}$$

第1成分という“1次元”
に3つの「評価基準」と
いう“3変数の情報”が
入っている!!!
第2成分 z_{i2} についても
同様に $k=2$ とした加重
和である

列側(評価基準)の3つの選択肢の合成変数!!!

$$j = 1, 2, 3$$

(つづき)

行の選択肢(レストラン)すべての第1成分スコアを求める.
もとの多次元情報がたしかに成分スコアに反映されている.

$$z_{i1} = \sum_{j=1}^3 l_{j1} x_{ij} = l_{11} x_{i1} + l_{21} x_{i2} + l_{31} x_{i3} = \frac{p_{i1}}{p_{i+} \sqrt{p_{+1}}} l_{11} + \frac{p_{i2}}{p_{i+} \sqrt{p_{+2}}} l_{21} + \frac{p_{i3}}{p_{i+} \sqrt{p_{+3}}} l_{31}$$

$i = 1, 2, \dots, 10$

↓

$$z_{11} = \frac{p_{11}}{p_{1+} \sqrt{p_{+1}}} l_{11} + \frac{p_{12}}{p_{1+} \sqrt{p_{+2}}} l_{21} + \frac{p_{13}}{p_{1+} \sqrt{p_{+3}}} l_{31}$$

$$z_{21} = \frac{p_{21}}{p_{2+} \sqrt{p_{+1}}} l_{11} + \frac{p_{22}}{p_{2+} \sqrt{p_{+2}}} l_{21} + \frac{p_{23}}{p_{2+} \sqrt{p_{+3}}} l_{31}$$

.....

.....

$$z_{10,1} = \frac{p_{10,1}}{p_{10+} \sqrt{p_{+1}}} l_{11} + \frac{p_{10,2}}{p_{10+} \sqrt{p_{+2}}} l_{21} + \frac{p_{10,3}}{p_{10+} \sqrt{p_{+3}}} l_{31}$$

いりふね

かりや

...

...

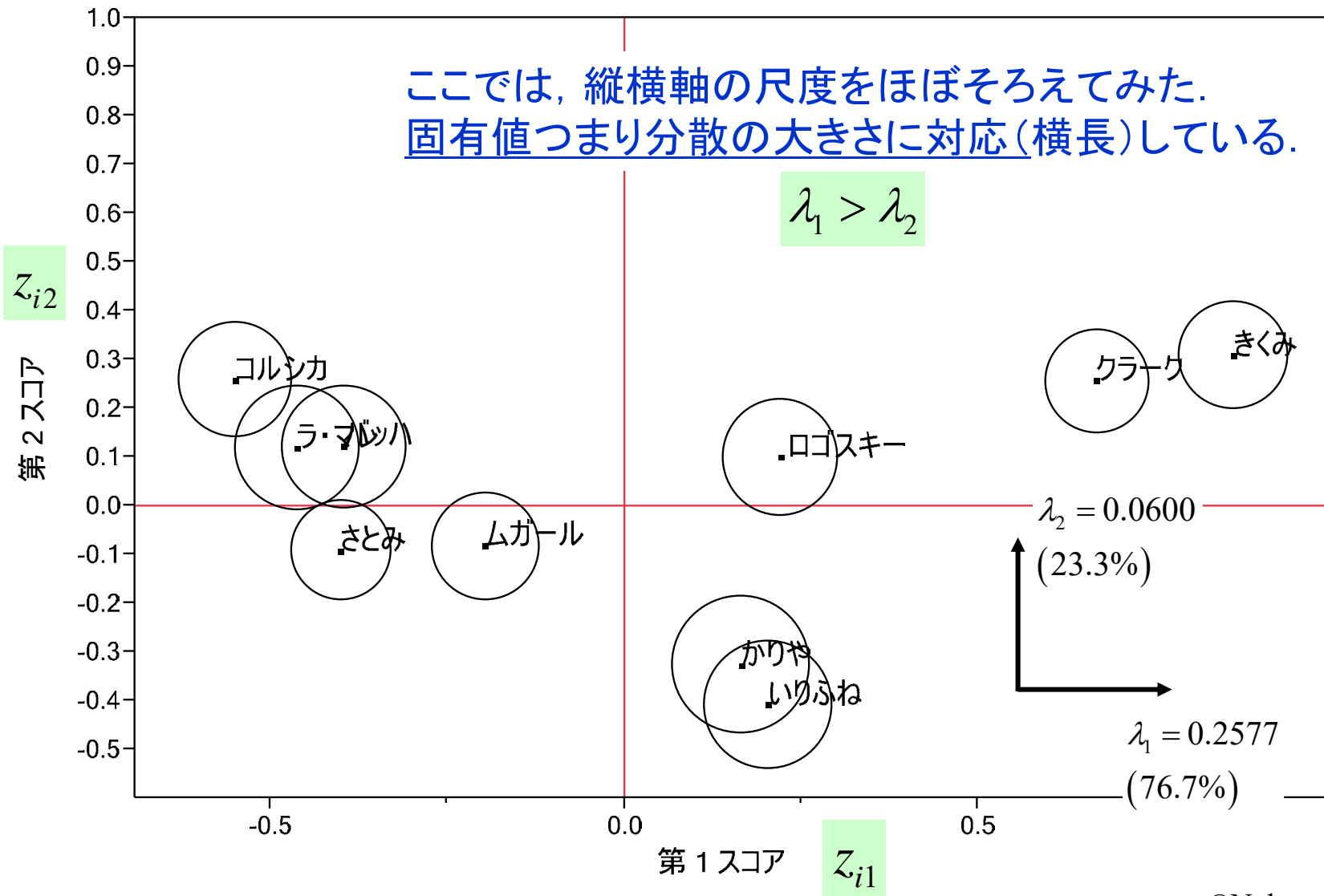
...

...

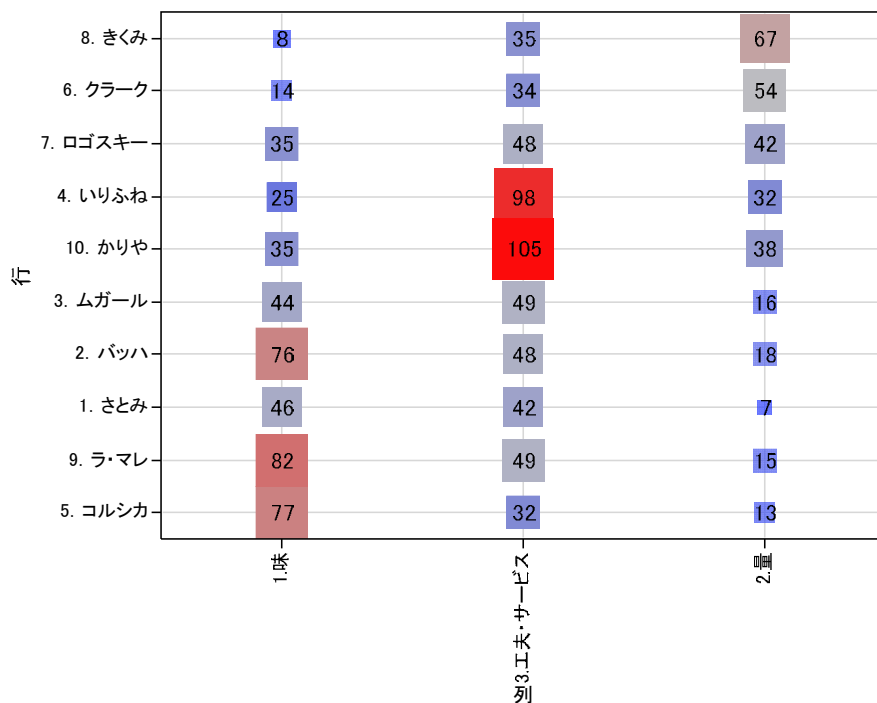
ロゴスキー

11に対応している

行成分スコア(レストラン)の観察(布置図)

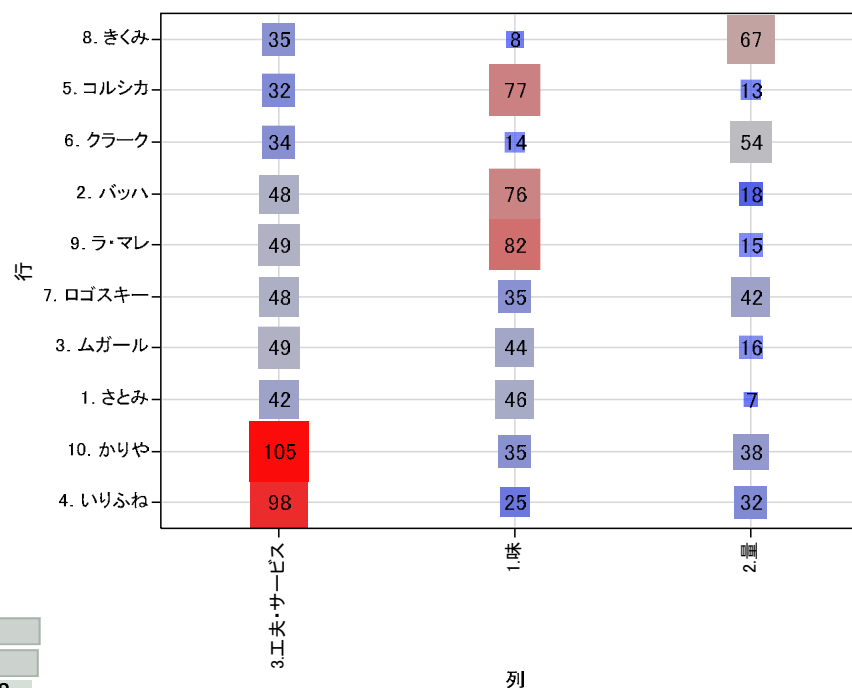


第1, 第2成分スコアによる行・列の並べ替え



第1成分スコアの並べ替え

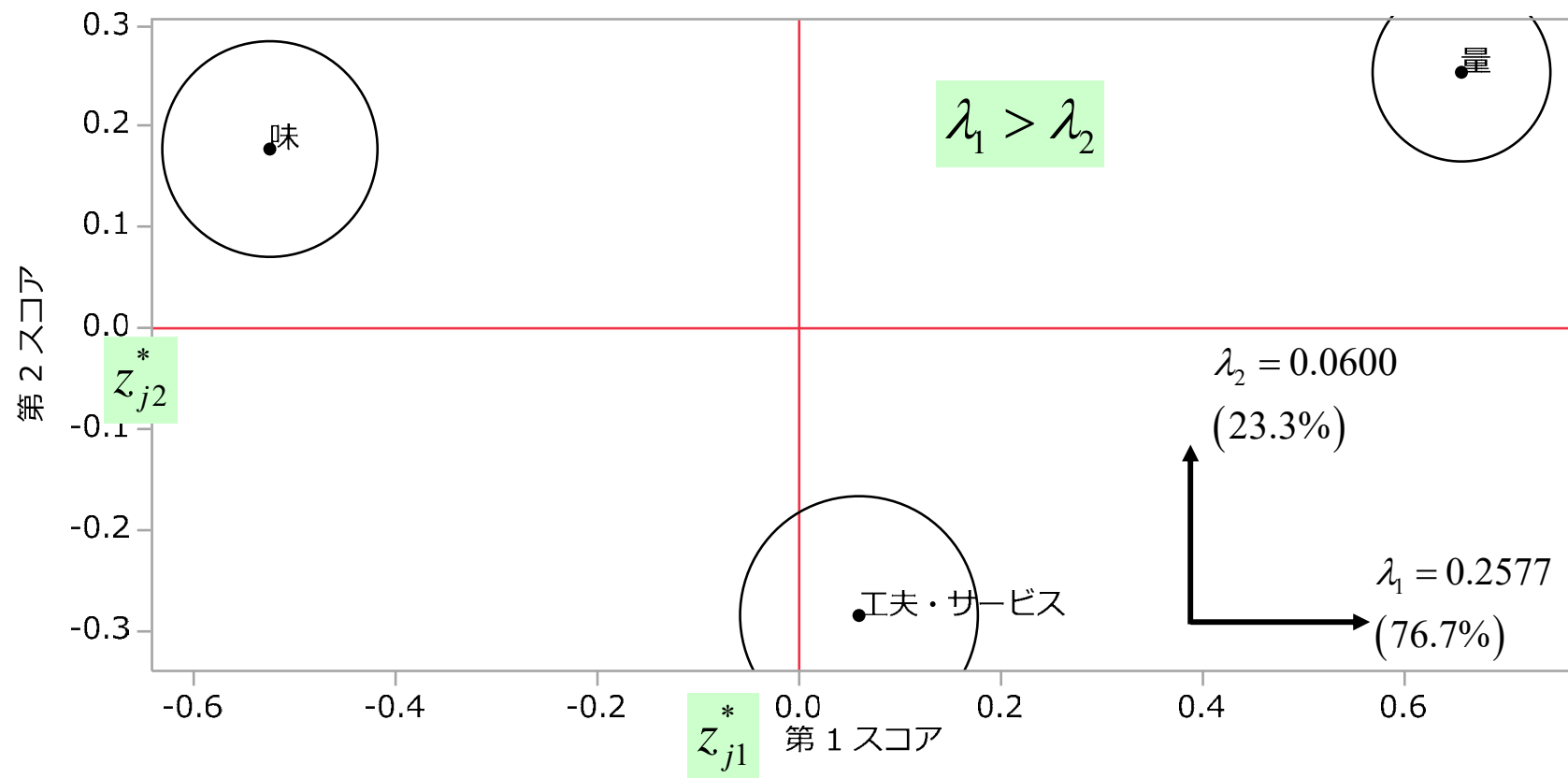
第2成分スコアの並べ替え



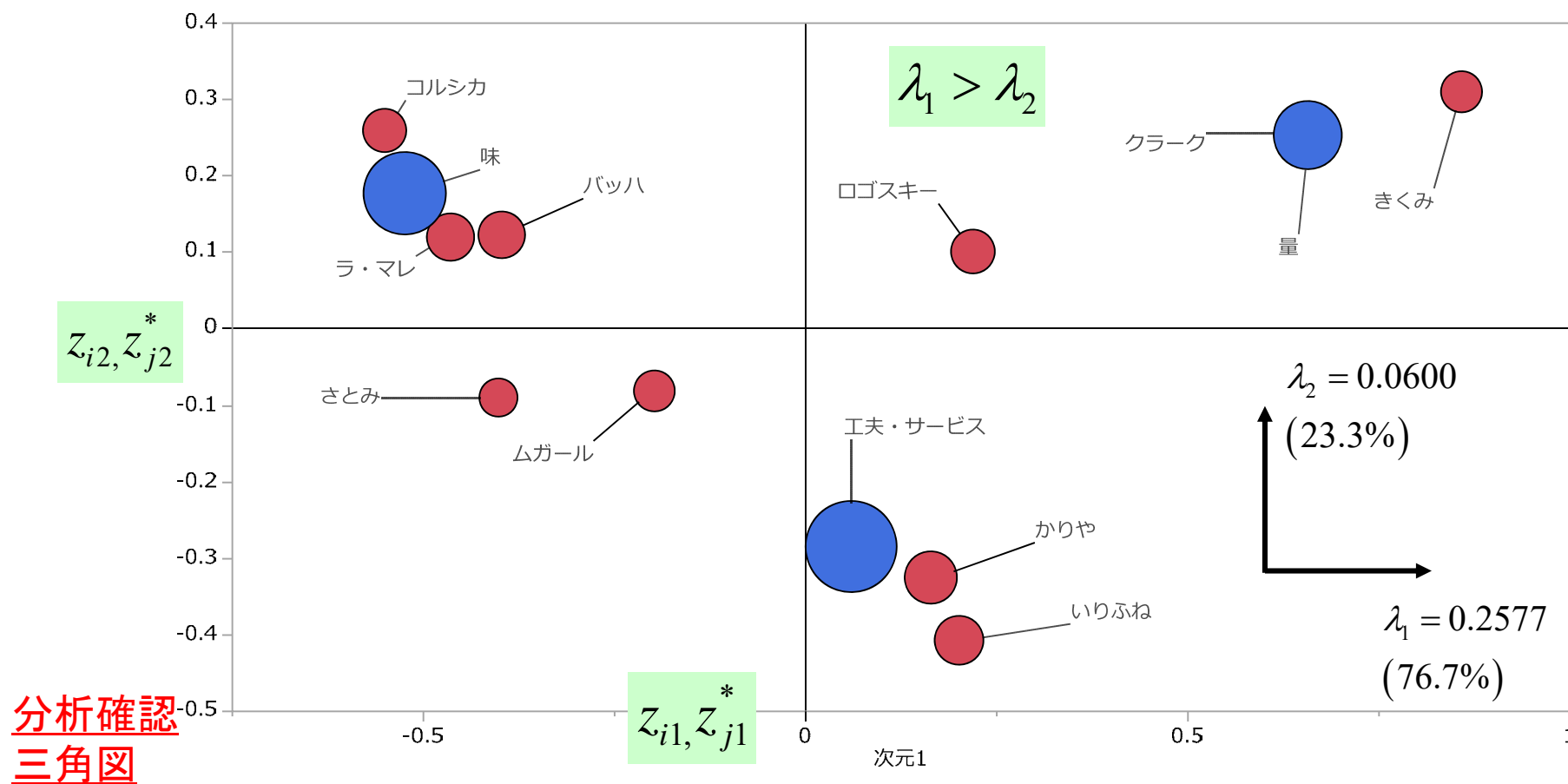
対応分析 特異値・固有値												
結果												
次元	特異値	固有値	割合(%)	.2	.4	.6	.8	累積(%)	.2	.4	.6	.8
1	0.44459	0.19766	76.7	<div><div></div></div>				76.7	<div><div></div></div>			
2	0.24498	0.06001	23.3	<div><div></div></div>				100	<div><div></div></div>			
固有値の合計 = 0.257678804157674												

● 特異値が成分スコア間の相関係数

列成分スコア(評価基準)の観察(布置図)



同時布置図(レストランvs評価基準)



分析確認
三角図

- これも第1成分の固有値 > 第2成分の固有値として描いた図.
- なるべく標準化しないほうがよいだろう.
- 見栄えではない, まず実寸で観察する.

布置図, 同時布置図の観察(要点)

- 成分スコアの布置図・同時布置図による“**多次元データ情報の視覚化**”は, 対応分析法の大きな特徴の1つ.
- 解釈をこれだけに頼ることには**リスク**がある.
- 2元データ表の寸法が小さいとき, たとえばクロス表のような場合には, えられる固有値数が少ないので寄与率が固有値が大きいはじめのほうの成分で説明がつくことが多い(情報のロスが少ない).
- (寸法が小さい)クロス表で少数次元で説明が難しいようなときは“データ表の作成方法”, “質問文の設計, 作り方”に問題があると考えたほうがよい.

(つづき)

- 「わからない」「どちらでもない」「ふつう」「無回答」といった選択肢がある場合は“はずれ値”となりやすい(尺度化の視点からいうと序列化がむずかしい).
- この時, 見かけ上大きい固有値が現れるが意味がない.
- 一方, 寸法の大きいデータ表では, たとえばテキスト型データで扱うようなデータ表では“大きな固有値は期待できない”ことが多い. [テキストにいくつか例を挙げた]
- しかし, 行, 列の関係を測っていることには変わりはない. 独立性の検定で有意でなくても, 対応分析では意味がある結果がみえる(ことが多い).

(つづき)

- 成分数(固有値の数)をいくつまで考えて分析すればよいかの判断がある. しかし, 決まった法則はない.
- よくある“累積寄与率が80%程度”は, 1つの目安にすぎない. 寸法が大きいデータ表では要注意.
- 多数の点が図中で重なり, 多数の成分スコアの点の観察もむずかしくなる. 視覚化の限界. 別の指標(寄与度).
- “はずれ値”的な成分スコアも増える.
- 布置図の観察のポイントの1つは, 図の周辺から観察することである. (後述)寄与度に関係.
- 理由は“中心あたり(原点, 重心周辺)は平均的な”パターンである(成分スコアの性質を思い出す). ただし, 多次元情報であることに注意.

(つづき)

- さらに“同時布置図”の観察には注意が必要である. この例について考える.
 - Q1: 評価「味」の周りに集まる3つのレストラン「コルシカ」「バッハ」「ラ・マレ」は「味」に距離が近い, といってよいか.
 - Q2: 「ロゴスキー」は3つの「評価基準」のどれからも遠い, といってよいか.
- 答えは「否」である. 正確に解釈するには“双対性”という性質を知ること. (後述)
- 成分スコアを, 行側からみた列側, あるいはその逆にみたときの関係がどうなっているか, である.
- 繰り返すが“対称性”が前提にあるので, 2元データ表の行と列との間に“方向性がある”因果性を考えるような場合は要注意である.

成分スコア観察の注意点(1)

- (1) まず、個々の成分スコアを“1次元”に観察する。もっとも情報が多い(最大固有値)、行と列との第1成分スコアを数直線上に並べて描いてみる。とくに、第1固有値の寄与率が高いときにはこの操作が大切である。

しかし“はずれ値”は、初めの方の成分、とくに第1成分に現れやすい。これは、はずれ値が成分スコアの分散を大きくするからである(原理からあきらか)。

- (2) つぎに、(固有値＝分散が大きい方から順に)2つの成分スコアに注目し、布置図を描き各点の布置の相対的な位置関係に注目する。たとえば、固有値の大きい方から、第1成分スコアと第2成分スコア、第3成分スコアなどを比べる。散布図行列などを用いるのもよい。

テキスト、I 部の45ページ、II 部の33～34ページあたり

成分スコア観察の注意点(2)

- (3) 場合に応じて、“絶対寄与度”をもとに成分軸を解釈する。
また、成分軸に解釈を与えるだけではなく、成分スコアの布置図の中での相対的な遠近、位置関係を観察する。
これらの観察には、後述の“寄与度”(相対寄与度、絶対寄与度)を目安とする。

重要度には濃淡がある。図は一部の成分であり情報量は同じではない。

- (4) “多重クロス表”(バート表)から求めたサンプルの成分スコアの解釈は「もとの変量・項目の選択枝のスコア」(つまりアイテム・カテゴリー型に展開した延べのカテゴリーに付与の成分スコア)であるから意味理解に注意する(とくに選択枝の並び順、順序関係に注意)。

◎ここからは、後述することの予告的な情報。先に示しておく。

成分スコア観察の注意点(3)

- (5) 多重クロス表から出発した場合の、固有値、寄与率の解釈は、通常の寄与率だけでは、大きくなることはほとんどないので注意する(高い寄与率が現れることがないことが数理的にわかっている)。
- (6) このとき、別の寄与率(調整済み寄与率)もあるので、それも参考にする。
- (7) 質問文の選択肢が順序尺度の場合には図中の選択肢の並び順に注意する。並び順が崩れたときには、その質問文の選択肢の作り方を再吟味する。

成分スコア観察の注意点(4)

- (8) この意味で成分スコアを用いたクラスター化操作には十分な注意が必要である。単純な k -平均法や階層的分類ではうまく対応できないことがある。カイ二乗統計量(つまり総変動)の分解を利用したクラスター化、つまり成分スコアによるクラスター化の工夫が必要である。
- (9) 布置図の上では、端のほうにある点から観察する。そして、どの成分軸に近いかを観察する。寄与度と併用。
- (10) 前述のように、“はずれ値”の存在に注意する。はずれ値はもとのデータ表の中の頻度分布の不均衡つまりプロフィールの不均衡から生じる。これは対応分析の特徴でもある。

成分スコア観察の注意点(5)

(11) 行と列の成分スコアの同時布置を考えたとき, それらの標準化(平均値=0, 分散=1とすること)の有無に注意する. それぞれを標準偏差で(特異値 あるいは固有値の平方根で)「標準化する場合」と「標準化しない場合」がある(表20). 通常の対応分析法では, いずれも標準化しないことが多い(組合せの「その1」; 成分スコアの分散は固有値 のまま).

表 20 成分スコアの分散の組み合わせ

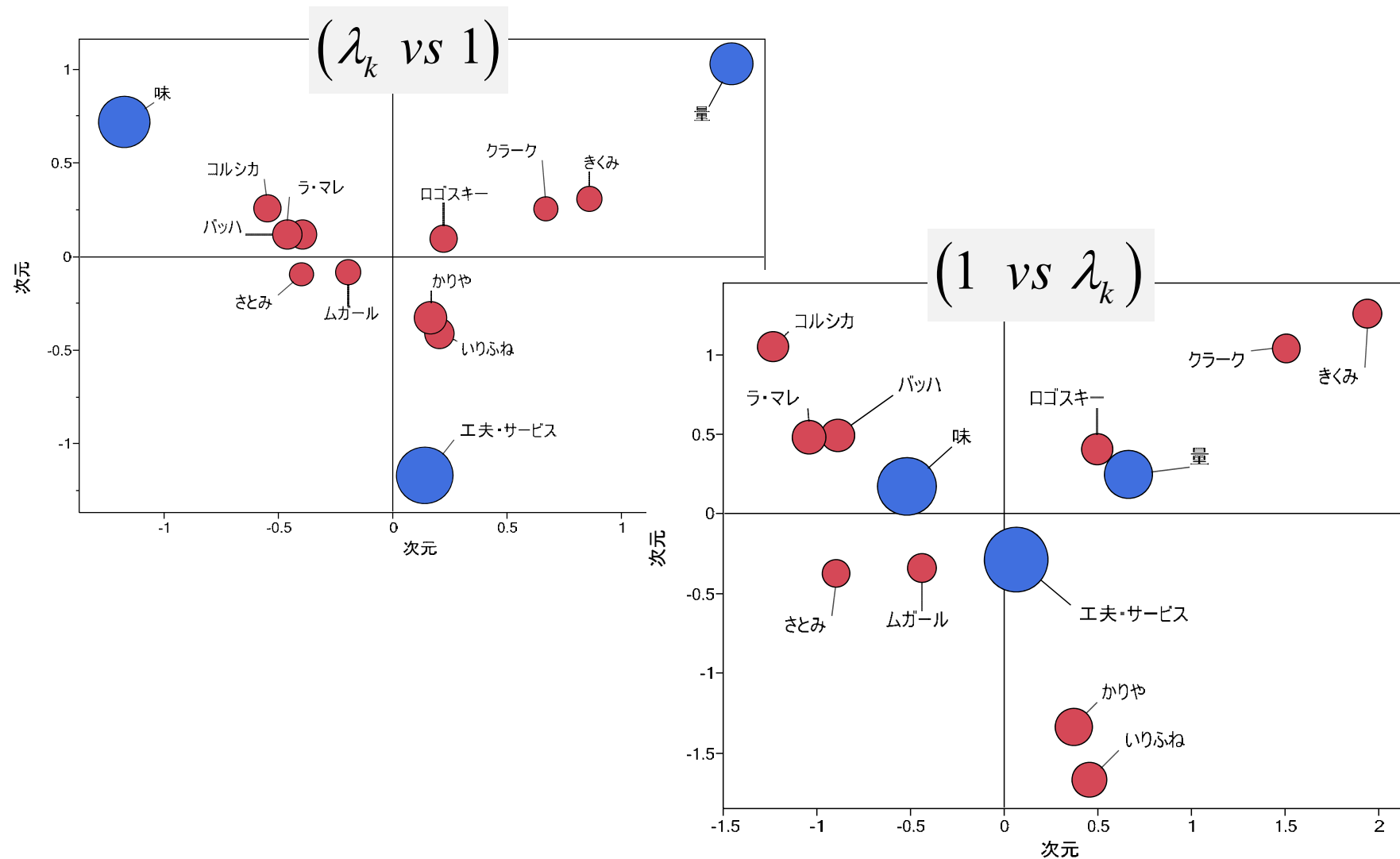
	組合せ	項目 I の選択肢の 成分スコア: z_{ik}	項目 J の選択肢の 成分スコア: z_{jk}^*
分散の大きさ	その 1	λ_k	λ_k
	その 2	λ_k	1
	その 3	1	λ_k
	その 4	1	1

成分スコア観察の注意点(6)

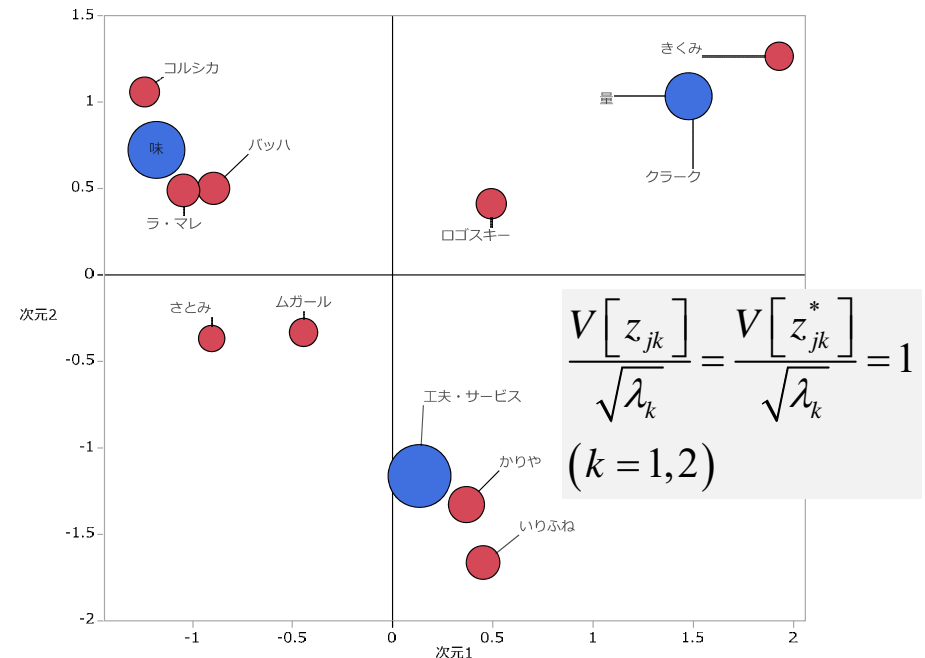
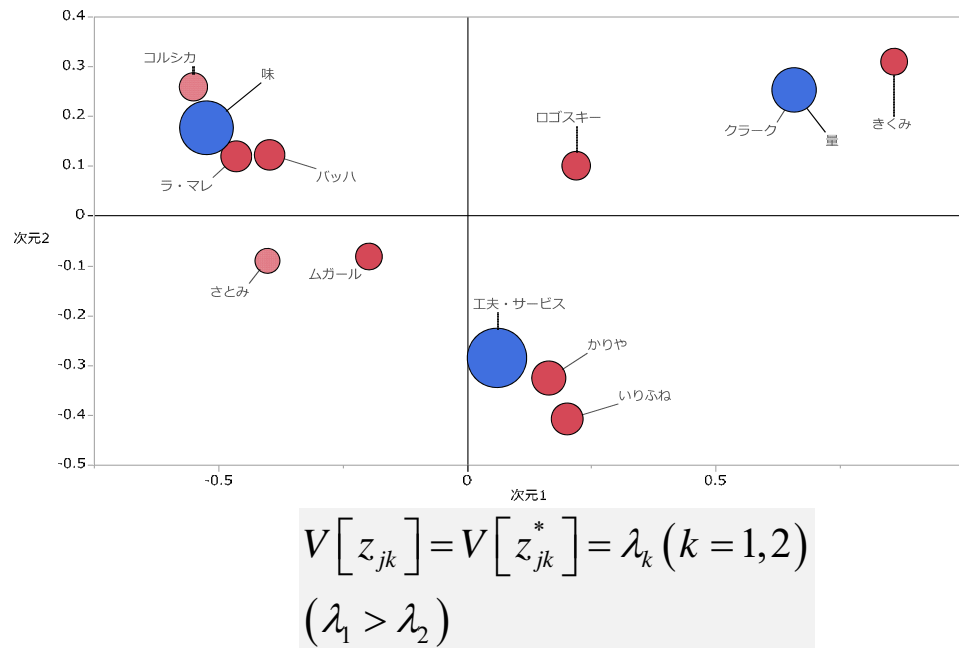
- 成分スコアを“標準化”するか, しないかで, 同時布置図の描き方は(少なくとも)4通りある. [標準偏差つまり特異値で割る]
- (フランス流の)対応分析法では「その1」の, 成分スコアの分散=固有値, を使う. 標準化をしない.
- 数量化法III類では「その4」となることが多い, つまりどちらも標準化することが多い(尺度が変わるので注意).
- 同時布置図の描き方についてのいろいろ議論がある.
- 配布テキストの中に, この尺度を変えた例を挙げた.
- いずれも正しいが, 意味が違うことに注意する.

標準化とは, 平均値=0, 分散/標準偏差=1と尺度変換すること.
テキストの I 部, 46ページあたり.

「その2」(上), 「その3」(下)の組合せで描いた図



「その1」(左), 「その4」(右)の組合せの図



- 左図は、一般に対応分析で出力する図。横軸は第1固有値，縦軸が第2固有値で固有値(分散)の大きさが異なる。
- 右図は、横軸，縦軸ともに分散＝1に標準化したとき。寸法が同じなので，成分スコアの大小判別ができない。

“双対性”について

- 双対性は次の“推移公式”で説明される性質のこと。
テキスト, 第 I 部, 41～42ページ

$$z_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^n \left(\frac{p_{ij}}{p_{i+}} \right) z_{jk}^* \quad (i \in I; k = 1, 2, \dots, K)$$

(“たすきがけ”になっている)

$$z_{jk}^* = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^m \left(\frac{p_{ij}}{p_{+j}} \right) z_{ik} \quad (j \in J; k = 1, 2, \dots, K)$$

双対性(duality)

“推移公式”(transition equation)を“遷移方程式”などともいう。

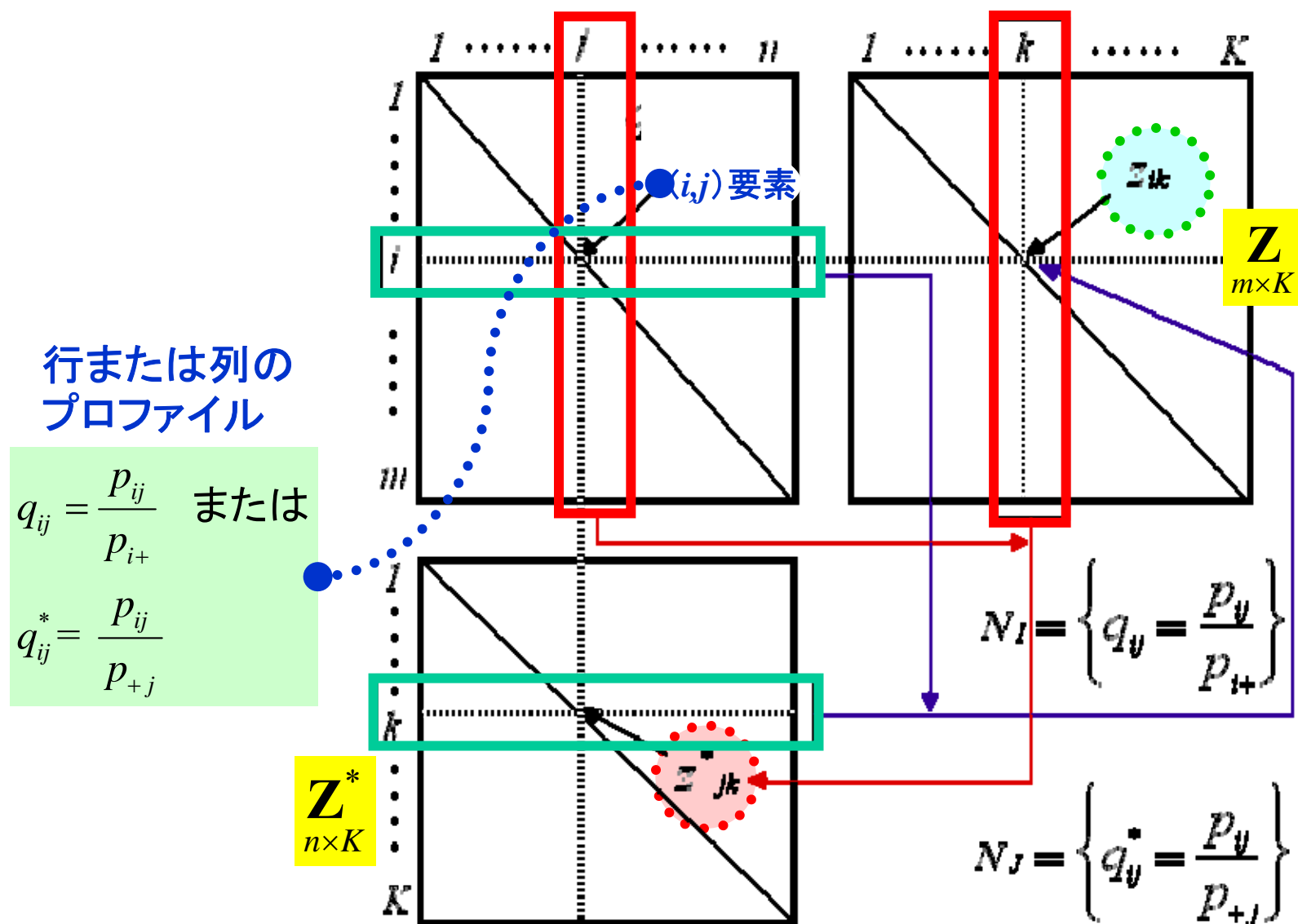
(つづき)

- “双対性”は布置図，とくに同時布置図を解釈するときに重要な働きをする.

[性質]

- 行成分スコアは，列成分スコアを重みとした，行プロフィールの加重平均となること.
- 列成分スコアは，行成分スコアを重みとした，列プロフィールの加重平均となること.
- 行成分スコアと列成分スコアの同時布置図は，この関係の中で，解釈すること.

双対性の模式図



“寄与度”（絶対寄与度と相対寄与度）

- “寄与度”とは、成分スコアの特徴を読み取る指標である。
- “絶対寄与度”と“相対寄与度”とがある。
- 絶対寄与度とは、簡単にいえば“ある成分 k の軸の解釈”に用いる指標。ある点（選択肢）の各成分への寄与。
- 相対寄与度は、ある点（選択肢）が、各成分軸により、どの程度“近似されるか”（説明力があるか）を示す指標。
- 相対寄与度は“平方相関”ともいう。
- この他「〇〇寄与度」と名付けた指標がいくつかあるが、ここでは上の2つを説明する。
- 式で説明はするが、うしろの分析例で確認するのがよい。

寄与度 (contribution), 絶対寄与度 (absolute contribution),
相対寄与度 (relative contribution) ・平方相関 (squared correlation)

絶対寄与度

第 k 成分における選択肢 $i \in I$ の絶対寄与度

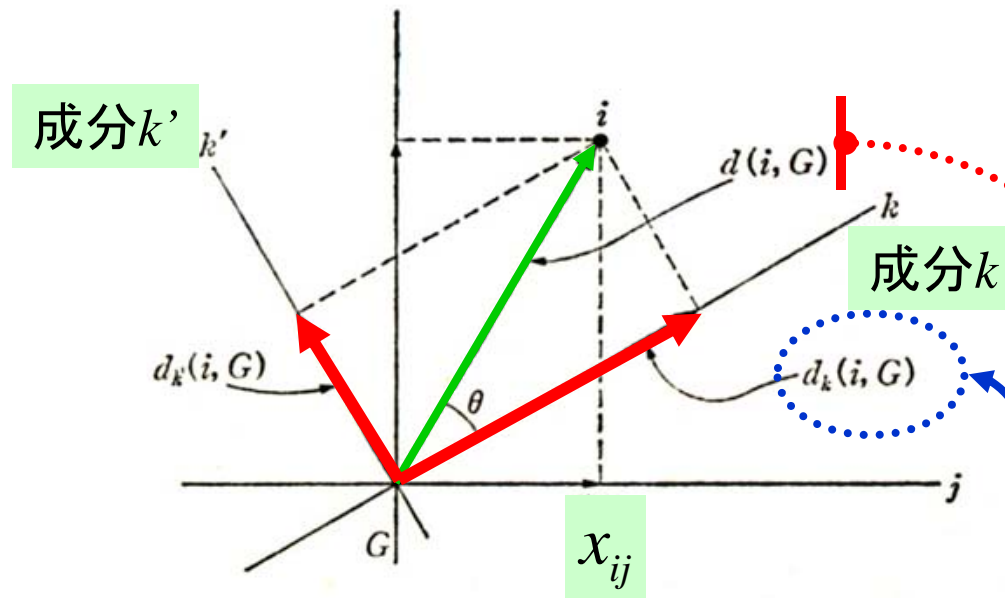
$$C_k(i) = \frac{p_{i+}(z_{ik})^2}{\lambda_k} \quad \left(\begin{array}{l} i \in I, k = 1, 2, \dots, K \\ K = \min\{m, n\} - 1 \end{array} \right), \quad \sum_{i=1}^m C_k(i) = 1$$

第 k 成分における選択肢 $j \in J$ の絶対寄与度

$$C_k(j) = \frac{p_{+j}(z_{jk}^*)^2}{\lambda_k} \quad \left(\begin{array}{l} j \in J, k = 1, 2, \dots, K \\ K = \min\{m, n\} - 1 \end{array} \right), \quad \sum_{j=1}^n C_k(j) = 1$$

- 例題でみるのがわかりやすいだろう。選択肢 i または j についての和
=100として%で使う

相对寄与度(平方相関)



- 選択肢 $i \in I$ からみたイメージ図.
- データの座標(重心座標系)と成分スコアの座標の関係に注目

$$d^2(i, G) = \sum_{k=1}^K d_k^2(i, G)$$

選択肢 $i \in I$ に対する相対寄与度

$$C_k^*(i) = \frac{d_k^2(i, G)}{d^2(i, G)} = \frac{z_{ik}^2}{\sum_{j=1}^n \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - p_{+j} \right)^2} \quad \begin{cases} i \in I, k = 1, 2, \dots, K \\ K = \min\{m, n\} - 1 \end{cases}$$

の見方を変えた
 入の定理
 回転図を想起

選択枝 $i \in I$ (の
 から重心 G まで

分母: $\sum_{j=1}^n (x_{ij} - \bar{x}_j)^2$

情報の見方を変えた
ピタゴラスの定理
前の回転図を想起

● 選択肢 $i \in I$ (のプロファイル)
から重心 G までの距離

(つづき)

- 同じようにして, “選択枝 $j \in J$ に対する相対寄与度”を以下のように約束する.
- 相対寄与度の解釈はやや面倒である. 基本はある点(選択枝)からみた, その選択枝のある成分への関係の度合いの強さ(近似の程度)のような意味をもつ. 成分軸の説明力.

選択枝 $j \in J$ に対する相対寄与度

$$C_k^*(j) = \frac{d_k^2(j, G)}{d^2(j, G)} = \frac{(z_{jk}^*)^2}{\sum_{i=1}^m \frac{1}{p_{i+}} \left(\frac{p_{ij}}{p_{+j}} - p_{i+} \right)^2} \quad \left(\begin{array}{l} j \in J, k = 1, 2, \dots, K \\ K = \min\{m, n\} - 1 \end{array} \right)$$

$$\sum_{i=1}^K C_k^*(i) = C_k^*(j) = 1 \quad (\text{成分} k \text{ についての和} = 100 \text{ として \% で使う})$$

レストラン分析の例：絶対寄与度を確認

< 行の絶対寄与度(%) >

レストラン	周辺確率(%)	成分1	成分2
いりふね	12.1	2.48	33.52
かりや	13.9	1.90	24.56
きくみ	8.6	31.98	13.64
さとみ	7.4	6.02	1.02
クラーク	7.9	17.89	8.66
コルシカ	9.5	14.53	10.58
バッハ	11.1	8.80	2.74
ムガール	8.5	1.66	0.95
ラ・マレ	11.4	12.36	2.69
ログスキー	9.7	2.38	1.63
	和	100.00	100.00



「ムガール」「ログスキー」はどこに位置するか？
「きくみ」は, 「いりふね」は？

レストラン分析の例：絶対寄与度を確認

＜列の絶対寄与度(%)＞

評価基準	周辺確率(%)	成分1	成分2
工夫・サービス	42.1	0.78	57.16
味	34.4	47.72	17.85
量	23.5	51.50	24.98
	和	100.00	100.00



- 「成分内」の方向に観察する. 列和＝100(%)となる.
- 数値の大きいほど, その成分を説明する力がある(“軸の解釈”に寄与).
- 「かりや」「いるふね」は2軸方向;「工夫・サービス」が2軸方向.
- 「きくみ」はどちらかというと1軸
- 「ムガール」「ロゴスキー」は, あまり成分に寄与せず(重心に近い, 平均的)

相対寄与度を確認

＜行の絶対寄与度(%)＞

レストラン	周辺確率(%)	成分1	成分2	和
いりふね	12.1	19.62	80.38	100.00
かりや	13.9	20.33	79.67	100.00
きくみ	8.6	88.53	11.47	100.00
さとみ	7.4	95.12	4.88	100.00
クラーク	7.9	87.18	12.82	100.00
コルシカ	9.5	81.88	18.12	100.00
バッハ	11.1	91.35	8.65	100.00
ムガール	8.5	85.18	14.82	100.00
ラ・マレ	11.4	93.81	6.19	100.00
ロゴスキー	9.7	82.78	17.22	100.00



相対寄与度を確認

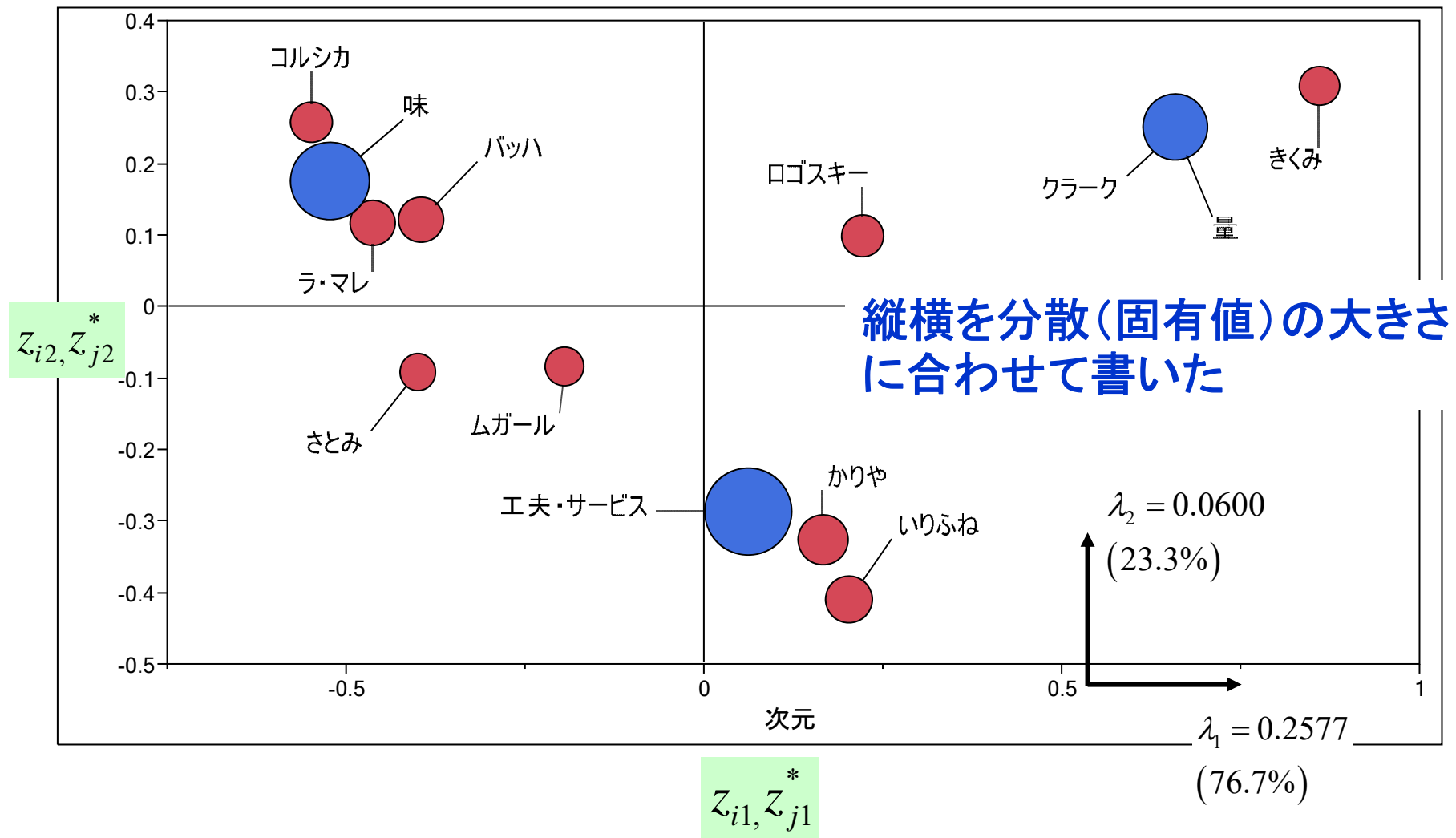
＜列の絶対寄与度(%)＞

評価基準	周辺確率(%)	成分1	成分2	和
工夫・サービス	42.1	4.30	95.70	100.00
味	34.4	89.80	10.20	100.00
量	23.5	87.16	12.84	100.00

こちら方向に観察

- ある点(選択肢)が「成分」を横断する方向で観察. 行和=100(%)となる.
- ある点(選択肢)が“どの成分で説明されるか”(よく近似しているか)を知る.
- 「いりふね」「かりや」は2軸の説明力が高い. 図の下の方に位置.
- その他のレストランは1軸での説明力が高い. 図の左右に分布.
- しかし, 値の2つの成分への振り分けには若干差違があるようだ.
- 「味」「量」は成分1, 「工夫・サービス」は成分2に説明力があり. 次元が異なる.

再確認：同時布置図（レストランvs評価基準）



“分布の同等性”

- カイ二乗距離を用いることで得られる特性の1つに“**分布の同等性**”がある. ⇔ストレッチ・プロファイルとする理由
 - 証明なしに数値例で確かめる.
- ① プロファイルが同じ列(つまり, 比率パターンが同じ列)を併合しても, 行間のカイ二乗距離は変化しない.
 - ② 同じように, プロファイルが同じ行(つまり, 比率パターンが同じ行)を併合しても, 列間のカイ二乗距離は変化しない.
 - ③ つまり, プロファイルが同じ行同士あるいは列同士を併合しても, 対応分析法の分析結果は変わらない.

分布の同等性 (distributional equivalency; equivalence of distribution)

数値例で確認しよう

回答者	銘柄A	銘柄B	銘柄C
回答者1	1	0	1
回答者2	1	0	1
回答者3	1	0	1
回答者4	1	0	1
回答者5	1	0	1
回答者6	0	1	0
回答者7	0	1	0
回答者8	0	1	0
回答者9	1	0	0
回答者10	1	0	0
回答者11	1	0	0
回答者12	0	1	1
回答者13	0	1	1
回答者14	0	1	1
回答者15	0	1	1
列和: f_{+j}	8	7	9

- アミカケ同じ色は“同じ回答パターン”
- 同じ行を積み重ねて併合圧縮したデータ表に作り替える.
- 第Ⅱ部, 39ページから.

同等のデータ表

回答者	銘柄A	銘柄B	銘柄C	行和: f_{i+}
回答者1～5	5	0	5	10
回答者6～8	0	3	0	3
回答者9～11	3	0	0	3
回答者12～15	0	4	4	8
列和: f_{+j}	8	7	9	24

- この2元データ表は前ページのデータ表と同じ情報.
- この点で, 主成分分析などと異なること.
- “カイ二乗距離”を使っていることでなり立つ関係.

得られた固有値・特異値と寄与率

表 21 得られた固有値と寄与率

固有値	特異値	寄与率
$\lambda_1 = 0.70273$	$\alpha_1 = 0.83829$	77.9%
$\lambda_2 = 0.19905$	$\alpha_2 = 0.44615$	22.1%

- 第Ⅱ部, 40ページから引用.
- 2つのどちらのデータ表から出発しても, “同じ”上の固有値, 寄与率ほかが得られる.
- テキストにある「表22」と「表23」を比べてみた.
- 列についての“同じプロファイルを併合”しても結果は変わらない. データの圧縮化.
- 参考: 行あるいは列のプロファイルが“類似する”ものは併合できる可能性がある. 自由回答やテキスト型データ.

“再生公式”とピアソンのカイ二乗統計量

- 対応分析法が与える情報は、ピアソンのカイ二乗統計量とどう関係するか.
- 全慣性(総変動)と、成分スコアの分散の総和(固有値の総和)の関係は述べた.
- ほかに“再生公式”がある. これは, “独立性の検定”と関連して非常に重要な関係式.
- 独立性の検定では「(クロス表の)行と列との関係が“ないとはいえない”」といった, 背理法的(二重否定的肯定)に「関係がありそう」と考えた.
- 再生公式では, 行と列との関係をより具体的に示す.

再生公式 (reconstitution formula)

$$p_{ij} = p_{i+} p_{+j} \left\{ 1 + \sum_{k=1}^K \frac{1}{\sqrt{\lambda_k}} z_{ik} z_{jk}^* \right\} = \underbrace{p_{i+} p_{+j}}_{\textcircled{1}} + p_{i+} p_{+j} \underbrace{\left\{ \sum_{k=1}^K \frac{1}{\sqrt{\lambda_k}} z_{ik} z_{jk}^* \right\}}_{\textcircled{2}}$$

($i \in I, j \in J, K = \min\{m, n\} - 1$)

$$\frac{p_{ij}}{p_{i+} p_{+j}} = 1 + \underbrace{\sum_{k=1}^K \frac{1}{\sqrt{\lambda_k}} z_{ik} z_{jk}^*}_{\text{独立モデルからのズレ}}$$

独立モデルからの乖離の程度!!!

$$f_{ij} = \left(\frac{f_{i+} f_{+j}}{N} \right) \left\{ 1 + \sum_{k=1}^K \frac{1}{\sqrt{\lambda_k}} z_{ik} z_{jk}^* \right\}$$

- 3つの書き方で示した.
- 第Ⅱ部, 36~37ページあたり.

(つづき)

- 式の右辺の第2項(②)を除外すると, ①だけ残る.
- これは, $p_{ij} = p_{i+}p_{+j}$ という“**独立モデル**”である.
- 第2項(②)には, 成分スコアの分散であり, 成分スコア間の相関係数(の二乗)である固有値が入っている.
- また, 行と列の成分スコアを含んでいる.
- つまり, この②は“独立モデルからの乖離度・ズレ”(gap, discrepancy)を表すと考えられる.
- こうなるためには, ストレッチ・プロフィールやそれに関連した(同等となる)“データ表”を考えることが必要.

$$\begin{array}{ccc}
 \underset{m \times n}{\mathbf{X}} = (x_{ij}) = \left(\frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} \right), & \underset{m \times n}{\mathbf{X}} = (x_{ij} - \bar{x}_j) = \left(\frac{p_{ij}}{p_{i+} \sqrt{p_{+j}}} - \sqrt{p_{+j}} \right) \Leftrightarrow \underset{m \times n}{\mathbf{Q}} = (y_{ij}) = \left(\frac{p_{ij}}{\sqrt{p_{i+} p_{+j}}} \right) \\
 \text{ストレッチ化} & \text{平均で中心化} & \text{対称化} \\
 \mathbf{Y}^* = (y_{ij}^*) = \left(\frac{p_{ij} - p_{i+} p_{+j}}{\sqrt{p_{i+} p_{+j}}} \right) & \text{(あるいはこれらを転置した行列)} &
 \end{array}$$

確認 & 演習: レストラン・データのデモをみる

- JMPには、2元クロス表の生成と、それに対応分析を行う処理の“標準機能”がある、
- このときに出発行列は「多変量構造型」つまり、「サンプル（回答者）×（項目）」のデータ表である。
- これとは別に、SAS/JMPで（セミナー用に）独自に作成したJMPスクリプトによる「対応分析」モジュールがある。
- 配布資料では、主にこのJMPスクリプトのモジュールを用いている。
- 両者を比べてみる。

例1: JMPの標準機能にある対応分析

例2: JMPスクリプトによる対応分析

分析例：情報に関する調査（ウェブ調査）

- 調査は2011年に実施.
- テキスト「第 I 部」, 5. 3. 3項にあげた例（75ページあたり）.
- 用いた質問文はテキストにある通り. マトリクス形式, チェックボックスによる複数選択による回答形式,
- 実は, これに類似の調査を「1996年」に「訪問留置・自記式」で行った.
- 「情報人間とは？」という課題.（林知己夫先生他）
- 約15年経過して, メディア環境もすっかり変わっている. そこで, 比較のために, 類似の調査を“ウェブ調査”で行った.

[調査の概要]

- 調査課題: 情報に関する調査
- 調査対象(標本抽出枠): あるウェブ・パネル(非公募型)に登録の首都40km圏に在住, 15歳以上69歳未満の男女(パネル構成の詳細情報は省く)
- 調査方式(モード): ウェブ調査
- 実施期間: 2011年9月9日(17時)~9月13日(10時)まで
- 計画標本数は766(人), 有効回収標本数は347(人), 参加率は45.3(%) [注: 参加率が高いとはいえない]

非公募型≡部分的に確率標本, ということ

用いた質問文の一部

Q17. 現在私たちは、情報を入手できる手段として数多くの情報源に囲まれており、それらの情報源についていろいろな意見が言われています。さて、以下でAからDの4つのことがあてはまる情報源にはどのようなものがあるでしょうか。あなたが「あてはまる」と思われるものをすべてお選びください。(それぞれいくつでも)

	A. 情報 が正 確	B. 情報 が詳 しい	C. 情報 量が多 い	D. 信頼 できる
	↓	↓	↓	↓
1. テレビの番組	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. ケーブルテレビ・衛星放送の番組	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. ラジオの番組	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. 新聞の記事(電子版を含む)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. 新聞の紙面広告(電子版を含む)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. 書籍(漫画・コミック以外)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. 一般の雑誌・週刊誌(漫画・コミック以外)の記事	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8. 各分野専門の情報誌の記事	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9. パンフレット・カタログ・ダイレクトメール	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10. 都・県や市・区など自治体の広報誌紙	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11. 所属する会や組織の会報・同人誌・ニュースレター	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

注意：
ストレートライニングなどは事前にチェックしよう。
応答度数表という2元データ表であってクロス表ではない。
行・列プロファイルが意味あるとして分析

調査後の分析で対応分析法
の適用を想定して設計。

回答者は23の「情報源」と9の
「評価項目」について、「あては
まる」場合を選ぶ。

	情報 が正 確	情報 が詳 しい	情報 量が多 い	信頼 できる
12. パソコンでみるインターネットサイト	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13. 携帯電話・PHS、スマートフォンでみるインターネットサイト	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14. インターネットブログ、ブログ	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15. ツイッター(Twitter)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16. 電子書籍(電子書籍端末や電子ブックリーダーで読む)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
17. ミクシィ(mixi)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
18. フェイスブック(Facebook)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
19. グリー(GREE)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
20. モバゲータウン	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
21. YouTube	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
22. ニコニコ動画	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
23. 1～22の中にはひとつもない	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

情報源(23選択肢)と評価項目(9選択肢)

情 報 源		評 価 項 目
1. テレビの番組	12. パソコンでみるインターネットサイト	情報が正確
2. ケーブルテレビ・衛星放送の番組	13. 携帯電話・PHS、スマートフォンでみるインターネットサイト	情報が詳しい
3. ラジオの番組	14. インターネットブログ、ブログ	情報量が多い
4. 新聞の記事(電子版を含む)	15. ツイッター(Twitter)	信頼できる
5. 新聞の紙面広告(電子版を含む)	16. 電子書籍(電子書籍端末や電子ブックリーダーで読む)	生活に欠かせない
6. 書籍(漫画・コミック以外)	17. ミクシィ(mixi)	役に立つ
7. 一般の雑誌・週刊誌(漫画・コミック以外)の記事	18. フェイスブック(Facebook)	世間の話題や流行を知る
8. 各分野専門の情報誌の記事	19. グリー(GREE)	商品を選び購入する
9. パンフレット・カタログ・ダイレクトメール	20. モバゲータウン	古くさい
10. 都・県や市・区など自治体の広報誌紙	21. YouTube	
11. 所属する会や組織の会報・同人誌・ニュースレター	22. ニコニコ動画	

$$m = 22, n = 9, K = \min\{22, 9\} - 1 = 8$$

(こういう大きさの2元データ表)

分析対象とする2元データ表

表 53 [情報源(23 選択肢)] × [評価項目(9 項目)]の2元データ表

質問項目	Q17_A 情報が正確	Q17_B- 情報が詳しい	Q17_C- 情報量が多い	Q17_D- 信頼できる
全サンプル数(回答者数)	347	347	347	347
1. テレビの番組	100	114	235	90
2. ケーブルテレビ・衛星放送の番組	43	91	103	44
3. ラジオの番組	65	68	86	54
4. 新聞の記事(電子版を含む)	140	153	131	131
5. 新聞の紙面広告(電子版を含む)	36	59	90	28
6. 書籍(漫画・コミック以外)	41	98	108	39
7. 一般の雑誌・週刊誌(漫画・コミック以外)の記事	19	84	139	12
8. 各分野専門の情報誌の記事	88	163	89	86
9. パンフレット・カタログ・ダイレクトメール	31	90	90	18
10. 都・県や市・区など自治体の広報誌紙	125	70	38	132
11. 所属する会や組織の会報・同人誌・ニュースレ	45	73	50	46
12. パソコンでみるインターネットサイト	26	118	274	24
13. 携帯電話・PHS、スマートフォンでみるインターネットサイト	18	63	166	15
14. インターネットブログ、ブログ	6	59	150	5
15. ツイッター(Twitter)	6	31	143	9
16. 電子書籍(電子書籍端末や電子ブックリーダー	22	39	103	19
17. ミクシィ(mixi)	9	36	128	11
18. フェイスブック(Facebook)	12	17	115	12
19. グリー(GREE)	6	20	99	5
20. モバゲータウン	8	20	100	4
21. YouTube	16	42	135	10
22. ニコニコ動画	7	22	122	6
23. 1～22の中にはひとつもない	54	23	11	62
99. 無回答	0	0	0	0

(つづき)

[情報源(23 選択肢)] × [評価項目(9 項目)]の2元データ表(つづき)

質問項目	Q18_A- 生活に欠か せない	Q18_B- 役に立つ	Q18_C- 世間の話題 や流行を知 る	Q18_D- 商品を選び 購入する	Q18_E- 古くさい
サンプル数	347	347	347	347	347
1. テレビの番組	226	166	248	51	20
2. ケーブルテレビ・衛星放送の番組	42	94	74	30	17
3. ラジオの番組	50	103	79	9	82
4. 新聞の記事(電子版を含む)	135	167	124	17	24
5. 新聞の紙面広告(電子版を含む)	29	62	74	59	16
6. 書籍(漫画・コミック以外)	57	100	71	33	16
7. 一般の雑誌・週刊誌(漫画・コミック以外)の記事	27	74	136	46	12
8. 各分野専門の情報誌の記事	24	137	63	47	10
9. パンフレット・カタログ・ダイレクトメール	12	76	59	155	29
10. 都・県や市・区など自治体の広報誌紙	41	148	25	7	79
11. 所属する会や組織の会報・同人誌・ニュースレ	11	102	25	5	67
12. パソコンでみるインターネットサイト	186	204	200	182	0
13. 携帯電話・PHS、スマートフォンでみるインターネットサイト	77	105	107	44	0
14. インターネットブログ、ブログ	34	67	135	14	1
15. ツイッター(Twitter)	21	43	117	5	1
16. 電子書籍(電子書籍端末や電子ブックリーダー	9	53	61	7	1
17. ミクシィ(mixi)	25	43	107	8	8
18. フェイスブック(Facebook)	14	35	94	7	9
19. グリー(GREE)	6	18	79	3	6
20. モバゲータウン	7	20	76	2	6
21. YouTube	27	88	120	4	1
22. ニコニコ動画	10	46	94	2	5
23. 1~22の中にはひとつもない	25	12	13	38	123
99. 無回答	0	0	0	1	0

ここで「22. この中にひとつもない」「99. 無回答」は除外した。
含めると、どうなるだろうか？

固有値, 特異値, 寄与率ほか

対応分析 特異値・固有値

結果

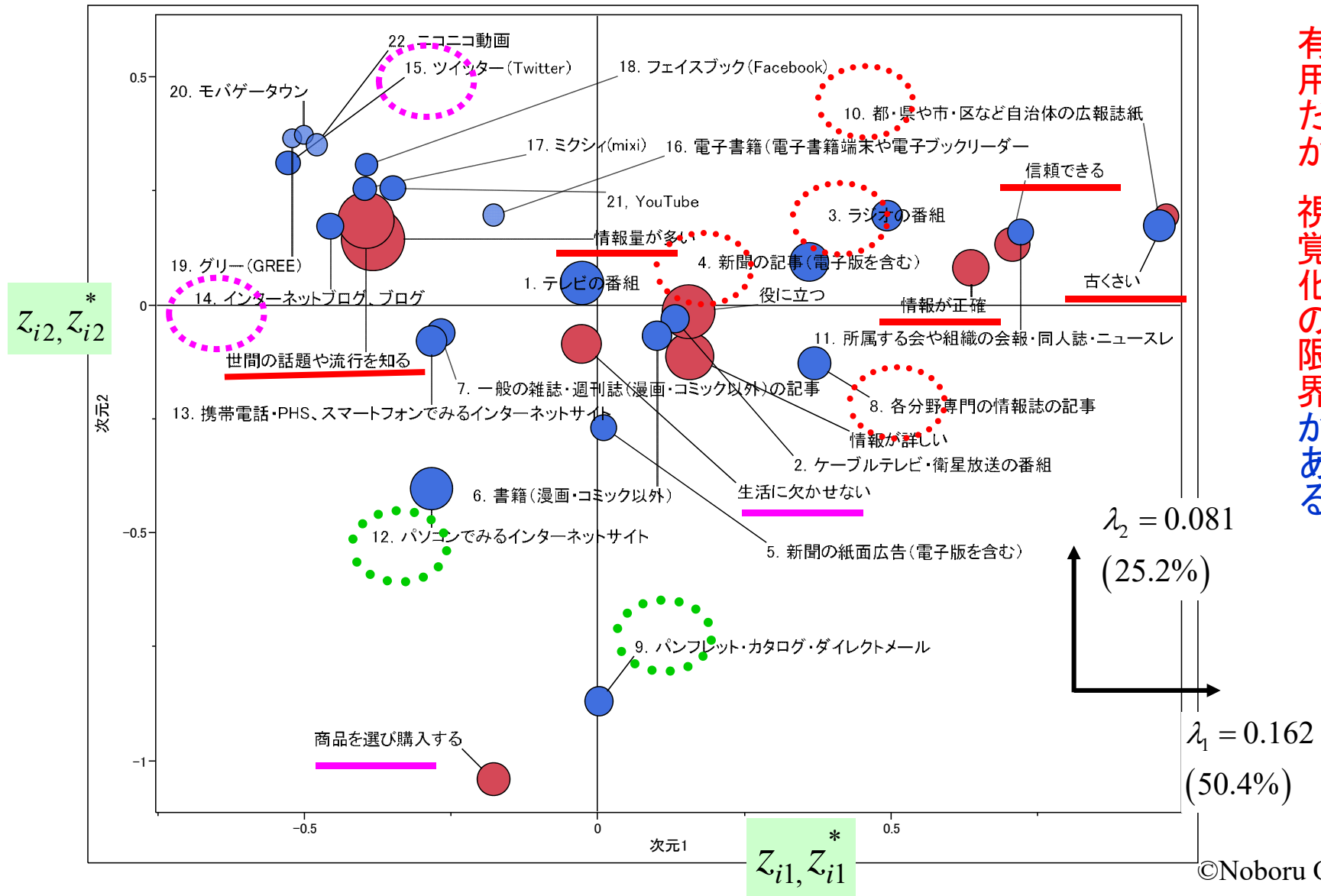
次元	特異値	固有値	割合(%)	.2	.4	.6	.8	累積(%)	.2	.4	.6	.8
1	0.40244	0.16195	50.4					50.4				
2	0.28439	0.08088	25.2					75.6				
3	0.20595	0.04242	13.2					88.8				
4	0.14903	0.02221	6.9					95.7				
5	0.08845	0.00782	2.4					98.2				
6	0.06512	0.00424	1.3					99.5				
7	0.03572	0.00128	0.4					99.9				
8	0.018	0.00032	0.1					100				

固有値の合計 = 0.321121971892488

$$\chi^2_p = N \times \sum_{k=1}^8 \lambda_k = 32272 \times 0.32112197 \dots \doteq 10363.24822 \quad (\text{ちなみにカイ二乗統計量はこうなる})$$

- 初めの2成分で, 寄与率は「約76%」となる.
- 3成分までで「88%」であるから, はじめの1~3成分の観察で十分だろう. $(n-1)=8$ 次元が縮退した.
- 特異値から, 初めの1~3成分の相関がやや高そう(このデータ表の寸法ならこの程度).

同時布置図の観察



この程度の点の数で、もはや判別はむずかしい？
有用だが、視覚化の限界がある。

絶対寄与度

情報源	周辺割合(%)	成分1	成分2	成分3	成分4
1. テレビの番組	10.2	0.05	0.30	25.52	10.37
2. ケーブルテレビ・衛星放送の番組	4.4	0.46	0.05	0.02	2.48
3. ラジオの番組	4.9	7.30	2.33	5.64	19.06
4. 新聞の記事(電子版を含む)	8.4	6.70	0.93	17.31	3.49
5. 新聞の紙面広告(電子版を含む)	3.7	0.00	3.28	0.96	0.08
6. 書籍(漫画・コミック以外)	4.6	0.29	0.26	0.52	1.15
7. 一般の雑誌・週刊誌(漫画・コミック以外)の記事	4.5	2.01	0.21	2.54	1.04
8. 各分野専門の情報誌の記事	5.8	4.84	1.14	0.11	41.47
9. パンフレット・カタログ・ダイレクトメール	4.6	0.00	42.73	15.90	0.01
10. 都・県や市・区など自治体の広報誌紙	5.4	30.76	2.08	0.29	0.71
11. 所属する会や組織の会報・同人誌・ニュースレター	3.5	11.10	1.11	13.09	3.29
12. パソコンでみるインターネットサイト	9.9	4.95	19.74	5.17	8.92
13. 携帯電話・PHS、スマートフォンでみるインターネットサイト	4.9	2.43	0.37	1.90	0.50
14. インターネットブログ、ブログ	3.9	4.97	1.43	0.18	0.45
15. ツイッター(Twitter)	3.1	5.32	3.69	0.57	0.15
16. 電子書籍(電子書籍端末や電子ブックリーダー	2.6	0.51	1.25	0.34	5.65
17. ミクシィ(mixi)	3.1	3.01	2.46	0.71	0.06
18. フェイスブック(Facebook)	2.6	2.49	3.03	1.78	0.30
19. グリー(GREE)	2	3.34	3.29	2.87	0.02
20. モバゲータウン	2	3.09	3.42	2.51	0.02
21. YouTube	3.6	2.75	2.96	0.04	0.72
22. ニコニコ動画	2.6	3.65	3.94	2.04	0.06
		100.00	100.00	100.00	100.00

第4成分までに登場
しない
第8成分までである



各成分(次元)の中で, どの「情報源」が(成分軸に)寄与するか.
各「情報源」の成分内での寄与の程度.

絶対寄与度

評価項目	周辺割合(%)	成分1	成分2	成分3	成分4
情報が正確	7.1	17.73	0.60	1.85	3.13
情報が詳しい	12.5	1.88	1.91	0.29	25.77
情報量が多い	22	20.03	5.69	4.23	0.14
信頼できる	6.5	20.16	1.45	2.74	3.98
生活に欠かせない	8.7	0.04	0.77	53.89	28.49
役に立つ	16	2.33	0.03	0.00	0.23
世間の話題や流行を知る	17.7	17.16	7.55	0.55	0.37
商品を選び購入する	6	1.19	80.37	3.54	0.63
古くさい	3.4	19.47	1.60	32.90	37.25
		100.00	100.00	100.00	100.00



各成分(次元)の中で, どの「評価項目」が寄与するか.
 各「評価項目」の成分内での寄与の程度.
 選択肢数が増えると探索のコツが必要になるかも.
 例: テキスト型データで「単語群」を扱うとき, かなりの数.

「絶対寄与度」は, 軸の解釈に役に立つ.

相対寄与度



情報源	周辺割合(%)	成分1	成分2	成分3	成分4	4成分まで和
1. テレビの番組	10.2	0.56	1.73	76.94	16.36	95.60
2. ケーブルテレビ・衛星放送の番組	4.4	46.75	2.53	0.57	34.57	84.41
3. ラジオの番組	4.9	56.88	9.08	11.50	20.36	97.83
4. 新聞の記事(電子版を含む)	8.4	54.12	3.75	36.64	3.87	98.37
5. 新聞の紙面広告(電子版を含む)	3.7	0.09	75.00	11.50	0.48	87.07
6. 書籍(漫画・コミック以外)	4.6	23.50	10.50	11.11	12.93	58.03
7. 一般の雑誌・週刊誌(漫画・コミック以外)の記事	4.5	60.06	3.06	19.86	4.27	87.25
8. 各分野専門の情報誌の記事	5.8	43.22	5.10	0.25	50.80	99.37
9. パンフレット・カタログ・ダイレクトメール	4.6	0.00	82.29	16.06	0.00	98.36
10. 都・県や市・区など自治体の広報誌紙	5.4	92.76	3.13	0.23	0.29	96.41
11. 所属する会や組織の会報・同人誌・ニュースレター	3.5	66.87	3.33	20.64	2.72	93.55
12. パソコンでみるインターネットサイト	9.9	27.81	55.36	7.61	6.87	97.65
13. 携帯電話・PHS、スマートフォンでみるインターネットサイト	4.9	64.53	4.88	13.22	1.84	84.47
14. インターネットブログ、ブログ	3.9	82.12	11.84	0.76	1.02	95.74
15. ツイッター(Twitter)	3.1	71.79	24.86	2.01	0.28	98.95
16. 電子書籍(電子書籍端末や電子ブックリーダー)	2.6	21.80	26.91	3.89	33.45	86.06
17. ミクシィ(mixi)	3.1	66.89	27.33	4.15	0.20	98.57
18. フェイスブック(Facebook)	2.6	51.14	31.14	9.59	0.83	92.70
19. グリー(GREE)	2	55.43	27.26	12.46	0.04	95.19
20. モバゲータウン	2	54.16	29.92	11.51	0.04	95.62
21. YouTube	3.6	54.67	29.41	0.23	1.98	86.28
22. ニコニコ動画	2.6	57.98	31.30	8.48	0.14	97.91

ある「情報源」は、どの成分での説明力が高いか(低いか)。

また、その「情報源」は何成分までで説明されるか(何成分まで考慮すべきか)。

例:「1. テレビの番組」は第3成分が近似がよい。

例:ソーシャル・メディア系は成分1, 2にまとまっている。

相対寄与度



評価項目	周辺割合(%)	成分1	成分2	成分3	成分4	4成分まで和
情報が正確	7.1	89.01	1.51	2.44	2.16	95.11
情報が詳しい	12.5	22.52	11.43	0.91	42.24	77.10
情報量が多い	22	81.58	11.57	4.51	0.08	97.74
信頼できる	6.5	87.33	3.15	3.11	2.36	95.94
生活に欠かせない	8.7	0.23	2.07	75.64	20.94	98.88
役に立つ	16	48.89	0.35	0.02	0.67	49.93
世間の話題や流行を知る	17.7	77.80	17.10	0.65	0.23	95.78
商品を選び購入する	6	2.79	93.95	2.17	0.20	99.12
古くさい	3.4	57.02	2.35	25.24	14.97	99.57

ある「評価項目」は、どの成分での説明力が高いか(低い)。

ここでは、多くの「評価項目」が、成分1で説明できそうである。

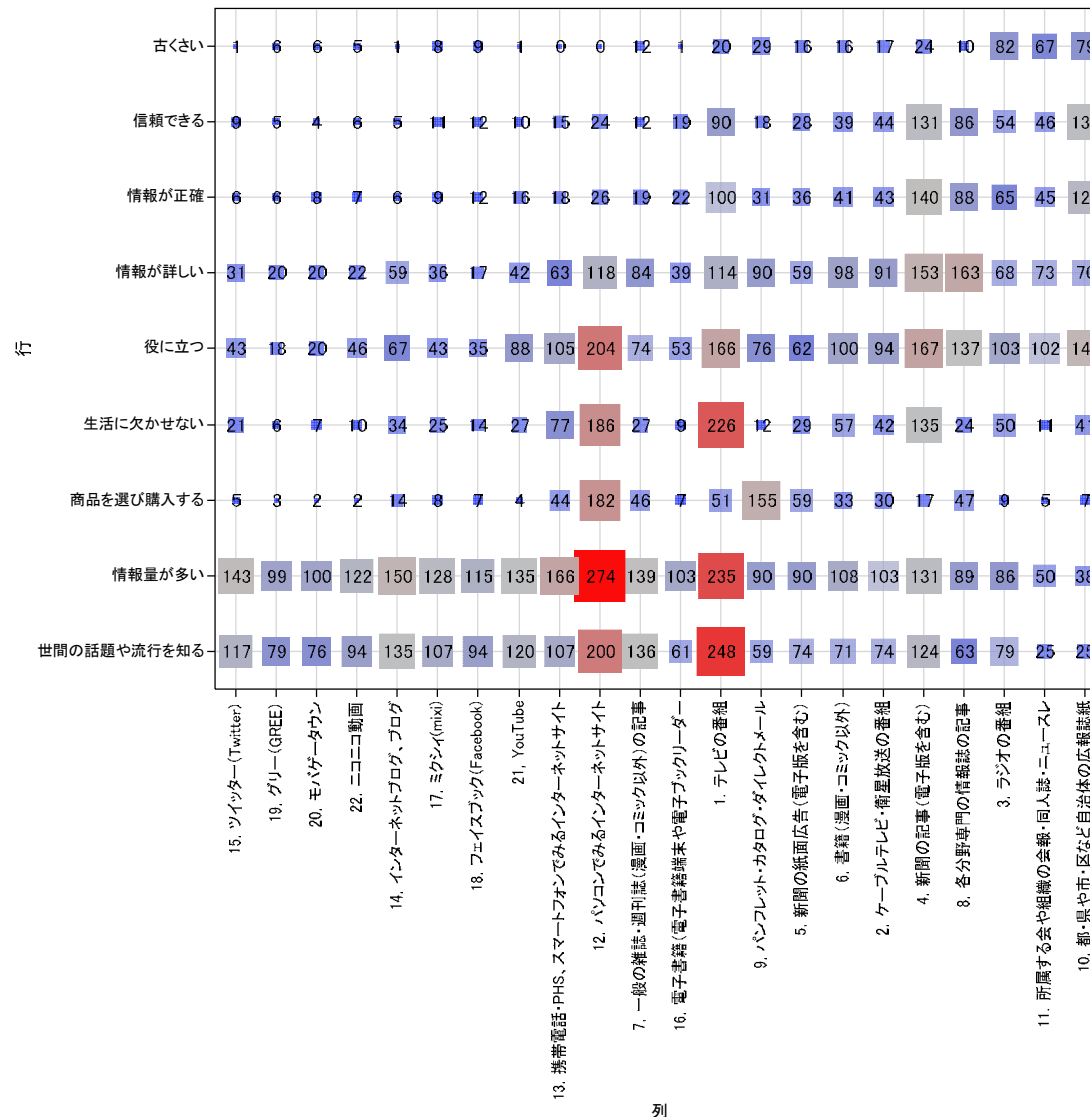
「情報が詳しい」「生活に欠かせない」などは、別の成分。(回答者に)他の項目とは違った意味に取られたのかもしれない。

「商品を選び購入する」は成分2で説明されるが、他からは分かれている。

「情報が詳しい」は、残り30%くらいが他の成分に分かれた。

「役に立つ」は、成分1である程度近似できるが(約49%)他に比べ十分でない。成分4までで情報が説明できないものがある、さらに高次の成分の観察が必要だろう。

参考: 対の散布図(第1成分スコアについて)



第1成分スコアについて
の2元データ表の行
と列の並べ替え

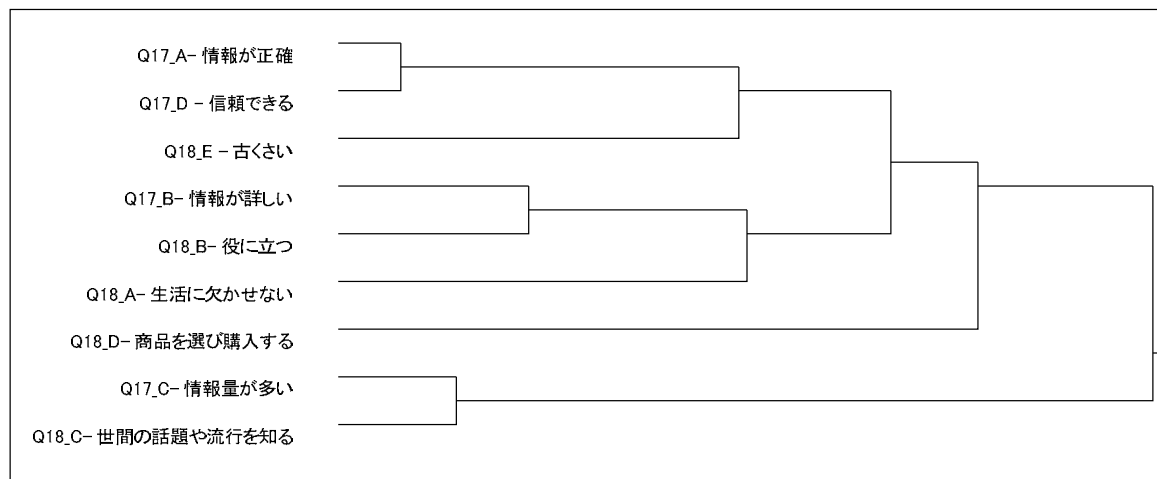
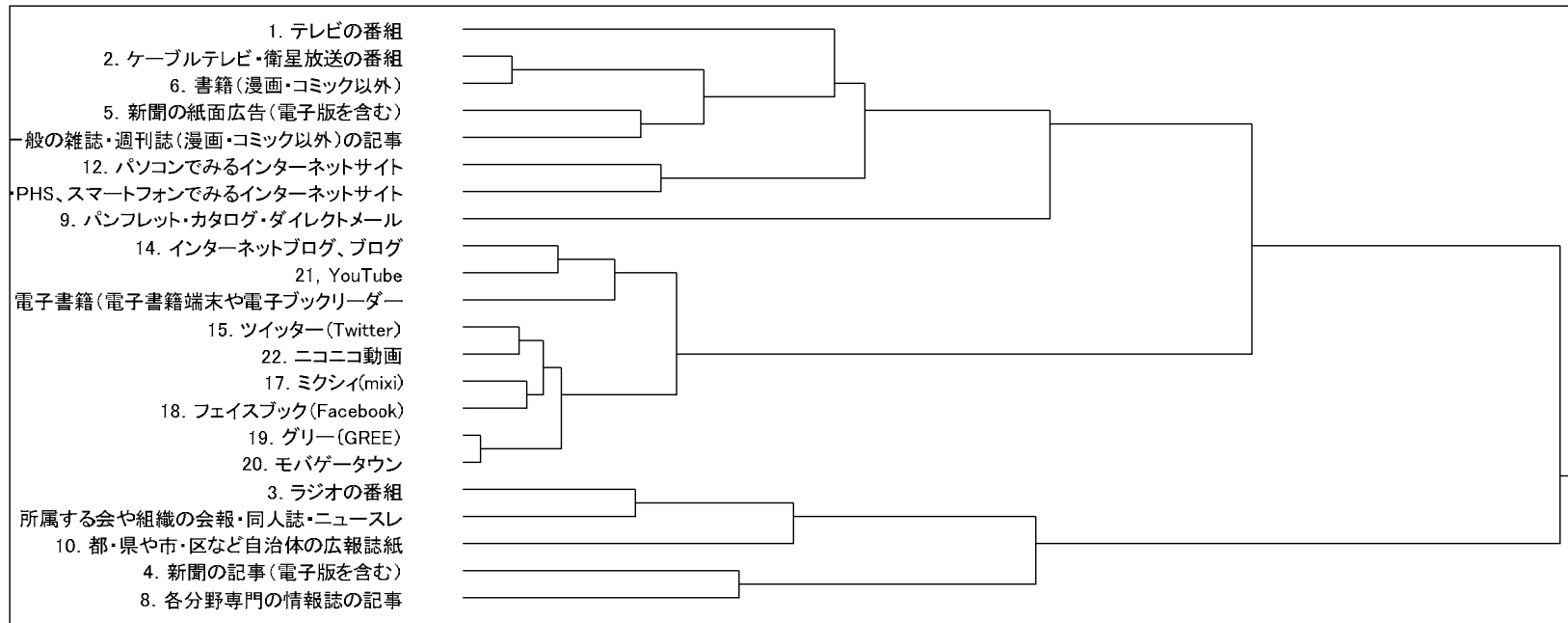
第1特異値=0.4024
(相関の程度)

度数(回答)の高いセル
の並び方に注目.
ほぼ対角に高い度数
が分布している.

参考：樹形図（デンドログラム）による観察

- クラスター化については、次回に述べる.
- もとのストレッチ・プロファイルのカイ二乗距離を用いることは、成分スコアのユークリッド距離を用いることに同じであった.
- 対応分析で得た成分スコアを用いたクラスター化を、行と列それぞれの成分スコアに対して、同等に、また同時に行えることを意味する.
- もとのクロス表のカイ二乗統計量の分解・圧縮化の操作を階層的に行うことに同じであることが分かっている.
- 行、列の選択肢数が少ない場合は、これで得た樹形図と成分スコア布置図とを比べることが有効である.
- 分類対象（つまり行、列の選択肢数）が増えると、視覚化に限界がある. ではどうするか？

「情報源」と「評価項目」の(同時)クラスター化



これはどのように得られるか？
行成分スコア, 列成分スコアを同時に(合わせて)クラスター化していないことに注意.

JMRAマーケティング・リサーチ講座

質的データのマイニング のための対応分析法

スライド資料[その4]

大隅 昇

ohsumi@ss.ij4u.or.jp

<http://wordminer.org/>

Copyright by Noboru Ohsumi

[その4]で述べること

- 予備知識として「分類手法の概要」を述べる.
- 階層的分類法と非階層的分類法の要点.
- ここで用いるクラスター化法の特徴, 手順の要点.
- 対応分析法におけるクラスター化法の基本.
- 対応分析法の基本操作[再確認]
- カイ二乗距離とユークリッド距離の関係[再確認]

(つづき)

- 成分スコアを用いた“ワード基準”によるクラスター化法の要点.
- “総変動”(全慣性)とクラスター化の関係.
- 生成クラスターの解釈と評価方法.
- 簡単な数値例, おもにレストラン・データで, これらの出力例を見ながら確認する.
- 最後に, 机上演習を兼ねた数値例を挙げる.

なぜ、クラスター化が必要か

- 対応分析に限らず，“分類操作”は，データ解析の必須手法である.
- 分けることで情報が集約・要約され，理解・解釈の見通しがよくなること. 探査的かつ発見的.
- 対応分析の場合，扱う初期の“2元データ表”の寸法 (m 行, n 列) が小さいとは 限らないこと.
- 調査におけるクロス表のような寸法が小さい例はむしろ稀，例外であること.
- よって，行情報 (m 個の要素)，列情報 (n 個の要素) の分類操作が必要であること.

(つづき)

- クロス表であっても、寸法(m 行, n 列)の“多次元データ”であること. この寸法情報が重要ということ.
- 説明に用いてきたクロス表でも、そうであったこと.
- よって、一般には“目視だけでは観察できない”情報であること.
- 換言すると、いわゆる“布置図”，“同時布置図”だけに頼る分析には限界があり、またリスクを伴うこと(誤解).

対応分析法とクラスター化法

- 対応分析法の“特性を活かしたクラスター化”というが、これは何を言っているのか。
- 非常にシステマティック、ロジカルに体系化されていること(巧みに仕組みが作られている)。
- 成分スコアに“**ワード基準**”を用いた分類が行える理由の1つは、成分スコアに“(平方)ユークリッド距離”を適用できるから。
- これを用いて、もとの2元データ表から得た行または列の成分スコアに対して分類を行う。
- しかし、成分スコアにそのままクラスター化手法を用いることはできないこと。統計ソフト利用上は注意。

(つづき)

- 2元データ表の行要素, 列要素の(両方向の)“**同時的分類を行うこと**”が可能. すでに簡単な例をみた.
- さらに“**カイ二乗統計量の分解**”(加法性)を利用して, クラスター化を行うことを考える.
- これはカイ二乗距離によるクラスター化に相当する.
- これらのクラスター化法は, 実は同じことを行っている.
- これらの要点を, なるべく数値例で読み解く.
- まず, 予備知識として, 分類手法を俯瞰する.

★メモ: WordMinerとJMPスクリプトで行うこと

- WordMinerでは, ここで述べる方法に加えて, いろいろな工夫がある.
- クラスタ化法として, 階層的分類法の典型手法である“ワード法”と, 非階層的分類法の典型手法の“ k -平均法”を使うこと. [ハイブリッド法という]
- 理由は, 度数が疎な行列の対応分析でえたボリュームが多く“クラスター(構造)”がはっきりしない, あるいははずれ値なども含む構造の成分スコアを効率的に分類したいため.
- JMPアドイン・スクリプトは, ここで述べる標準的な手順による成分スコアに, ワード基準を用いるクラスタ化を行う.

分類手法の概要(予備知識)

- “分類”は分野を問わず, ものごとの基本操作である.
- “分けることは, 分かること”と言われよう, 「分ける」という操作を通じて理解を深める(「分けて知る」こと).
- 生物分類や系統学の支流として“数値分類法”が登場.
 - 動物分類学・動物系統学(Systematic Zoology)ほか多数
- パターン認識では“教師なし分類”(unsupervised classification)がある. ⇔教師あり分類≡判別分析
- 社会心理, 社会科学系で, 因子の分類の視点から“クラスター分析”が登場.

分類(classification), 数値分類法(numerical taxonomy),
動物系統学(Systematic Zoology), 教師なし分類・判別分析,
クラスター分析(cluster analysis)

(つづき)

- データ解析分野では、あれこれ取り混ぜて“自動分類法”として研究が進展した.
- 分類手法は無数にある. 分野, 提唱者により, さまざまな名称が付与.
- 視点, 用途が多様で, 一意的にこれだと分けられない.
- “分類手法の分類”が必要になる理由(これも無数の議論がある).
- データ解析の視点からは, コンピュータ支援を前提に, 自動的に“分けて知る”操作のこと. よって“自動分類法”という. [当然, 自動化の限界がある]

自動分類法 (automatic classification), 分類手法の分類,
分類の目的と達成手順

(つづき)

- 「なぜ分類するのか」(分類の目的)と, それを「どう実現するか」(手順つまり算法・アルゴリズム)は分けて考えることが肝要.
- 原則, コンピュータ利用が前提となる(不可欠). プログラミング技術が鍵である.
- 一般に言われている“分類手法の分類”を概観する.
- 視点や用途が様々であり, 一意的にこれがクラスター化法だと言えない (†).
- “分類問題”は, 実は難問であり, 未解決の課題が多い.

(†)よくある“「クラスター分析」を行ったらこれこれとなった”という言い方は不適切ということ(説明が不十分).
また, さまざまな分類問題がある.

自動分類とクラスター化(ここでの約束)

- ここでは, 分析対象の“自動分類”で, なんらかの“群(グループ; group)”あるいは“クラスター(cluster)”を作る操作を“クラスター化”と呼ぶ.
- この用語の用い方, 別称について, さまざまな議論があるが, ここでは“クラスター”と“クラスター化”とする.
- “クラスター”の原義は「房状・塊状のもの」であるが, あとで簡単な数値例でみるように, “クラスターをどう定義するか”が重要な課題である. [未解決?]
- “クラスター化”とは, 性質の類似した仲間同士を自動的に分類してグループを作る“基準を決めること”と, それを実現する“アルゴリズム(算法)”のこと, と考える.

★メモ:「分類法の分類」の例(見方はさまざま)

階層的分類法 Hierarchical method	凝集型 (agglomerative)	ウォード法 (Ward's method), 群平均法, 単連結法 (最短距離法), 最長距離法, 重心法など無数 「組み合わせ的手法」でいくつかは統合化される
	分枝型 (divisive)	AID, THAIDなど, CART: 二進木解析, 関連分析 (association analysis) など
	グラフ理論や関係代数の応用 組合せ理論, 計算機幾何学	ファジィ・グラフ, 最小張り木 (MST: minimum spanning tree) DEMATEL法・ISM法
非階層的分類法 Non-hierarchical method	分割最適化型分類法 (Partitioning type)	k-平均法 (k-means) ファジィk-平均法
	多くの変形手法	ISODATA法 (iterative self-organizing data analysis-A) ダイナミック・クラスタリング

最小張り木 (MST) ⇔ 最短経路問題・巡回セールスマン問題 (難問の1つ)

“クラスター”とは

- “最適”あるいは“万能”の決まった手法があるわけではない.
- “クラスター”(cluster)とは何か, つまり“どのようなクラスターを考えるか”は“算法”(アルゴリズム)で異なる.
- “クラスター”は存在するものではなく, “ある基準のもとに作り出すもの, 生成するもの”と考えること.
- よって“クラスター化”という呼び方をしよう.
- “はずれ値”の影響を受けやすい. はずれ値に非常に敏感な方法と, 逆にはずれ値検出に鈍い方法と, 様々である.
- 手法による長所・短所がある. これは目的に応じて“使い分けること”を意味する. あとで簡単な例.

クラスタリング=clusteringだが, むしろ造語すると“clustered”(クラスター化する)
その他, clumping,, lumping, aggregating, grouping, …とさまざまな用語あり
はずれ値(outlier)

基本的な考え方

- 分類対象が“類似している(似ている)”を約束する指標や“目的関数(最適化基準)”が必要となる. [最適化基準]
⇒アルゴリズムとして記述
- 何をもって“似ているとするかの基準”の設定が必要. これを“クラスター化基準”という. [クラスター化基準]
- 似ている程度を測る“類似度”,あるいは似ていない程度を測る“非類似度”(例:距離)を用意することが必要.
- この最適化基準と類似度・非類似度の関係があまり明らかでない.
- 比較的分かっている実用的な基準が“分散または平方和”つまり“平方ユークリッド距離”を使うこと.

類似度, 非類似度, 距離

(つづき)

- クラスター化基準が異なれば, 異なるグループの作り方となる. 分類結果は何通りも作れることになる.
- どのようにクラスター化したかの“クラスター評価基準”が必要(示すこと). [クラスター評価基準]
- クラスター化基準とクラスター評価基準には曖昧性がある(どちらも似たようなことを言っている).
- 喩えれば「卵と鶏のどちらが先」式の議論.
- これが(生成した)“クラスターの解釈”をどう考えるかに関連する(どういう基準で作ったクラスターか).

最適化基準, クラスター化基準, クラスター評価基準

数値例による手法の特徴の確認

- 凝集型階層的分類法 (AHC: agglomerative hierarchical clustering methods) の典型的な算法である“**組み合わせ的手法**” (combinatorial method) で説明されるいくつかの手法がある.
- “**ワード法**” (Ward's method) もその1つとして表現される. しかし, ここではこの方式 (組み合わせ的) には従わない.
- (WordMinerではワード法に加えて) 非階層的分類法のうち, 分割化型分類法の代表的手法である“**k-平均法**” (k -means method) を用いる.

“組み合わせ的手法” (combinatorial method) [Lance & Williamsの方法]
ワード法, k -平均法ともに, クラスター化基準として“平方和や分散”を用いる.

2変量のトイ・データによる基本の確認

$n = 8, p = 2$ の2変量データ

個体番号	X1	X2
1	4	4
2	6	7
3	8	6
4	7	5
5	2	5
6	3	2
7	6	8
8	2	3

- 見通しをよくするため, 簡単なトイ・データを用意する.
- 寸法が, 個体数が8, 変量数が2のデータ表.
- これをいま, 下のよう書く.
- 必要な“統計量”を求める.

$$\mathbf{X}_{n \times p} = (x_{ij}) \begin{pmatrix} i = 1, 2, \dots, n \\ j = 1, 2, \dots, p \end{pmatrix}$$

\Downarrow

$$\mathbf{X}_{8 \times 2} = (x_{ij}) \begin{pmatrix} i = 1, 2, \dots, 8 \\ j = 1, 2 \end{pmatrix}$$

必要な統計量を求める

個体番号	X1	X2
1	4	4
2	6	7
3	8	6
4	7	5
5	2	5
6	3	2
7	6	8
8	2	3
平方和	S_{11}	S_{22}
分散	s_1^2	s_2^2

平均ベクトル(重心): $\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} = \begin{pmatrix} 4.75 \\ 5.00 \end{pmatrix}$

平方和積和行列: $\mathbf{S} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} 37.5 & 21.0 \\ 21.0 & 28.0 \end{pmatrix}$

平方和: $S_{11} = \sum_{i=1}^8 (x_{i1} - \bar{x}_1)^2$, $S_{22} = \sum_{i=1}^8 (x_{i2} - \bar{x}_2)^2$

積和: $S_{12} = S_{21} = \sum_{i=1}^8 (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$

- 平方和積和行列の対角要素が平方和, 非対角要素が積和(相関を測る項)

個体番号	X1	X2
1	4	4
2	6	7
3	8	6
4	7	5
5	2	5
6	3	2
7	6	8
8	2	3
平方和	S_{11}	S_{22}
分散	s_1^2	s_2^2

- 平方和積和行列Sの対角要素の和を跡和(トレース)いう, これが“平方和の和”となる.
- つまり, 平均(重心)からの平方距離の和(変動の総量).

$$tr(S) = S_{11} + S_{22}$$

$$= \sum_{i=1}^8 \left[(x_{i1} - \bar{x}_1)^2 + (x_{i2} - \bar{x}_2)^2 \right] = 37.5 + 28.0 = 65.5$$

- 2つの変量の分散(不偏分散)

$$s_1^2 = \frac{1}{n-1} S_{11} \Rightarrow s_1^2 = \frac{1}{8-1} \times 37.5 \doteq 5.36$$

$$s_2^2 = \frac{1}{n-1} S_{22} \Rightarrow s_2^2 = \frac{1}{8-1} \times 28.0 = 4.0$$

統計量を要約確認

統計量	X1	X2
平均	\bar{x}_1	\bar{x}_2
平方和	S_{11}	S_{22}
分散	s_1^2	s_2^2

 \Rightarrow

実現値	X1	X2
平均	4.75	5.00
平方和	37.5	28.0
分散	5.36	4.00

$$S_{11} + S_{22} = 65.5 \text{ (平方和の和)}$$

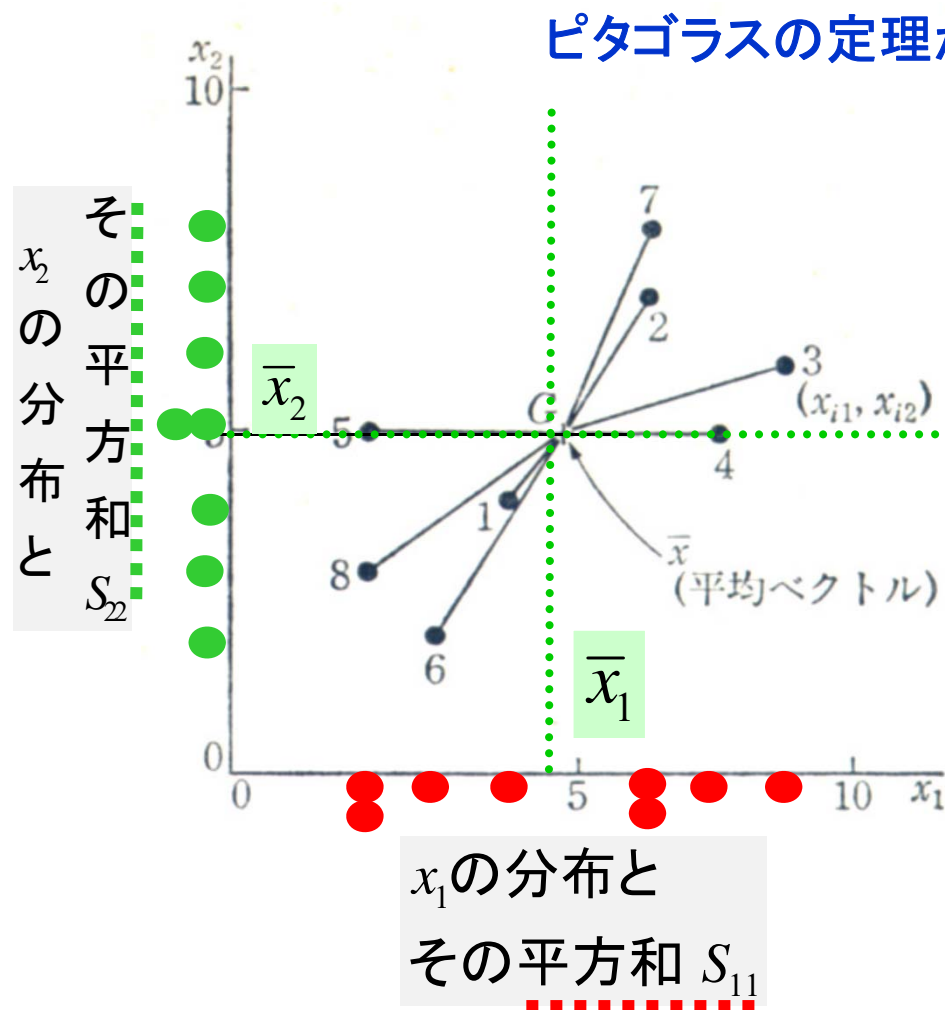
$$s_1^2 + s_2^2 = 9.36 \text{ (分散の和)}$$

- こうした統計量が算出できるということは“量的データ”であるから可能であること.
- 次に登場する“平方ユークリッド距離”も同様.

ここで調べること

- “平方和”, “分散”と距離, とくに“平方ユークリッド距離”はどのような関係にあるか.
- [2つの変量の平方和の和]
 = [(平均) 平方ユークリッド距離] つまり
 [2つの変量の分散の和]
 = [(平均) 平方ユークリッド距離の平均]
- 分散あるいは平方和を考えることは(平方)ユークリッド距離を考えることに同じ.
- 変量数は多変量になっても仕組みは同じである.
- 確認: 対応分析で得た成分スコアに平方ユークリッド距離を適用できるとしたことを思い出す.

平方和は, ...



ピタゴラスの定理から

(x_{i1}, x_{i2}) から重心までの平方距離

$$\sum_{i=1}^8 \left[\underbrace{(x_{i1} - \bar{x}_1)^2}_{x_1 \text{の平方和}} + \underbrace{(x_{i2} - \bar{x}_2)^2}_{x_2 \text{の平方和}} \right]$$

$$= S_{11} + S_{22} = \underline{65.5}$$

平方和積和行列Sの跡和=平方和の和

重心: G

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} = \begin{pmatrix} 4.75 \\ 5.00 \end{pmatrix}$$

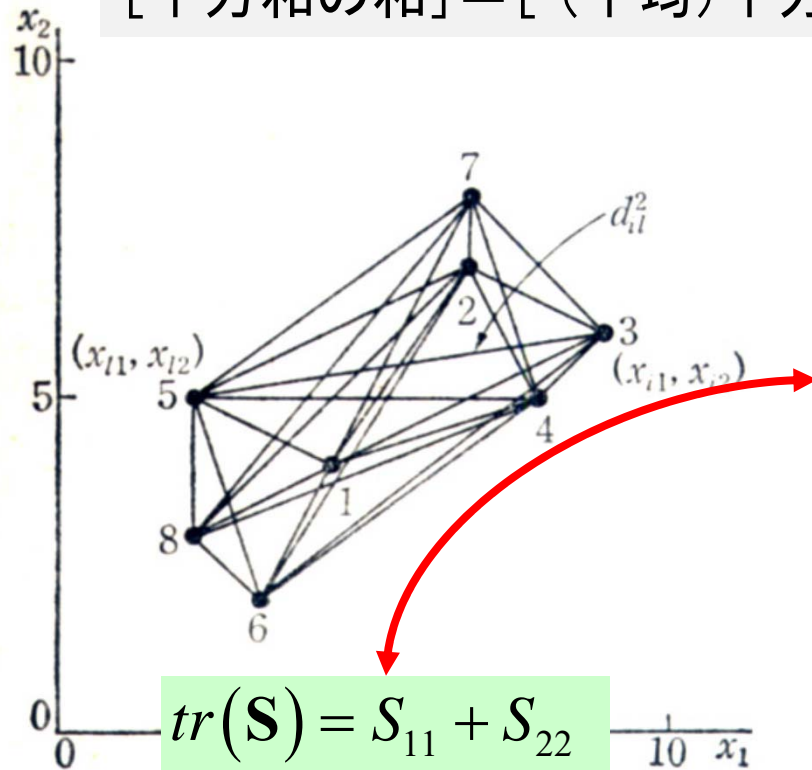
$$S = \begin{pmatrix} 37.5 & 21.0 \\ 21.0 & 28.0 \end{pmatrix}$$

(平方和積和行列)

平方ユークリッド距離は...

ここで確認すること

[平方和の和] = [(平均) 平方ユークリッド距離]



$$\left\{ \begin{array}{l} \frac{1}{n} \sum_{i=1}^n \sum_{l=1}^n d_{il}^2 \quad [(\text{平均}) \text{平方ユークリッド距離}] \\ \quad \quad \quad (i < l \text{ の和}) \end{array} \right.$$

$$\text{ここで, } d_{il}^2 = \underbrace{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2}_{\text{点}i\text{と点}l\text{との平方ユークリッド距離}}$$

↓ [全距離の和を作り平均]

$$\frac{1}{8} \sum_{i=1}^8 \sum_{i'=1}^8 d_{ii'}^2 = \frac{524}{8} = \underline{65.5} (= S_{11} + S_{22})$$

$$\left({}_8C_2 = \frac{8!}{(8-2)!2!} = 28 \text{通りの距離がある} \right)$$

$$tr(\mathbf{S}) = S_{11} + S_{22}$$

$$\frac{1}{n} \sum_{i=1}^n \sum_{l=1}^n d_{il}^2 = tr(\mathbf{S})$$

(i < l の和)

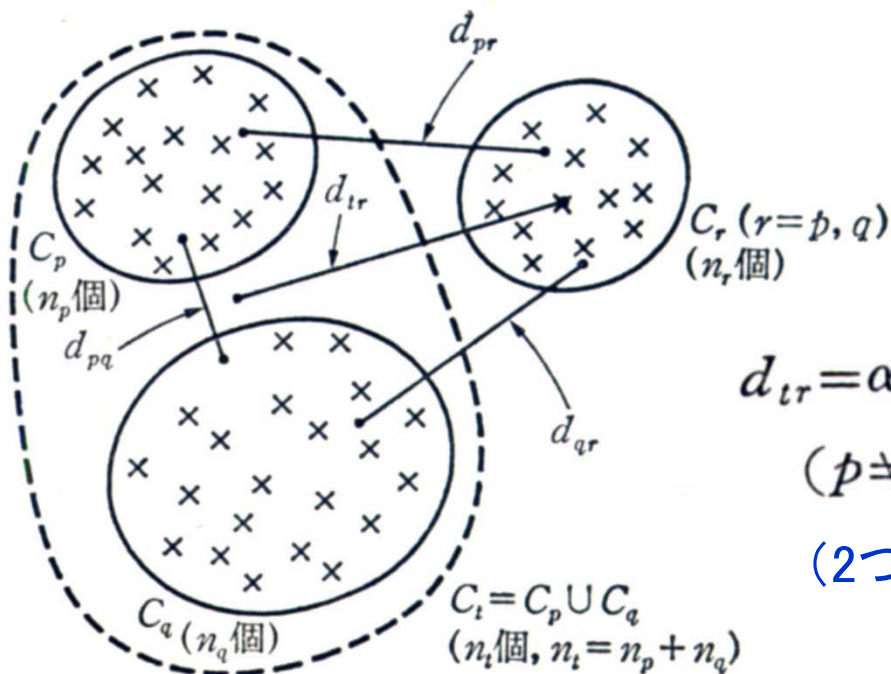
平方ユークリッド距離は「距離の公理」を満たさない

なぜワード法と k -平均法か

- 対応分析法で得た成分スコアは“数量”つまり“量的データ”として扱えること(なんども確認した).
- もとの2元データ表のプロフィール間の“カイ二乗距離”は、得られた成分スコア間の“ユークリッド距離”であること.
- “平方”カイ二乗距離と“平方”ユークリッド距離の関係も調べた.
- よって, “平方”ユークリッド距離を用いるクラスター化法として, ワード法や k -平均法を用いることは自然な対応.
- 直前に調べたように, 平方ユークリッド距離と平方和とは, 実は同じ情報である.

★参考: 組み合わせ的手法とは

- 基本は, Lance & Williams(1967, 1977)によって提案され, その後さまざまな手法が誕生した. 提案年次に注目.
- 基本的な階層的分類法がこの考え方で統一的に議論できることが特徴. 手法例は次ページ.



[クラスター間の距離の更新式]

$$d_{tr} = \alpha_p d_{pr} + \alpha_q d_{qr} + \beta d_{pq} + \gamma |d_{pr} - d_{qr}|$$

$$(p \neq q, r \neq p, q)$$

(2つのクラスターを併合するときの定義式)

★おもな組み合わせ的手法の例

手 法 名	α_p	α_q	β	γ	$\alpha_p + \alpha_q + \beta$	単調性
最短距離法	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	1	○
最長距離法	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	1	○
群平均法	$\frac{n_p}{n_t}$	$\frac{n_q}{n_t}$	0	0	1	○
加重平均法	$\frac{1}{2}$	$\frac{1}{2}$	0	0	1	○
重心法	$\frac{n_p}{n_t}$	$\frac{n_q}{n_t}$	$-\frac{n_p n_q}{n_t^2}$	0	$1 + \beta$	×
メジアン法	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0	$\frac{3}{4}$	×
可変法	$\frac{1}{2}(1 - \beta)$	$\frac{1}{2}(1 - \beta)$	$\beta < 1$	0	1	○
ワード法	$\frac{n_p + n_r}{n_t + n_r}$	$\frac{n_q + n_r}{n_t + n_r}$	$-\frac{n_r}{n_t + n_r}$	0	1	○

いろいろな手法があることを知っておく

大隅(1989)から.
これ以外にも多数の手法が統一的に記述できる.

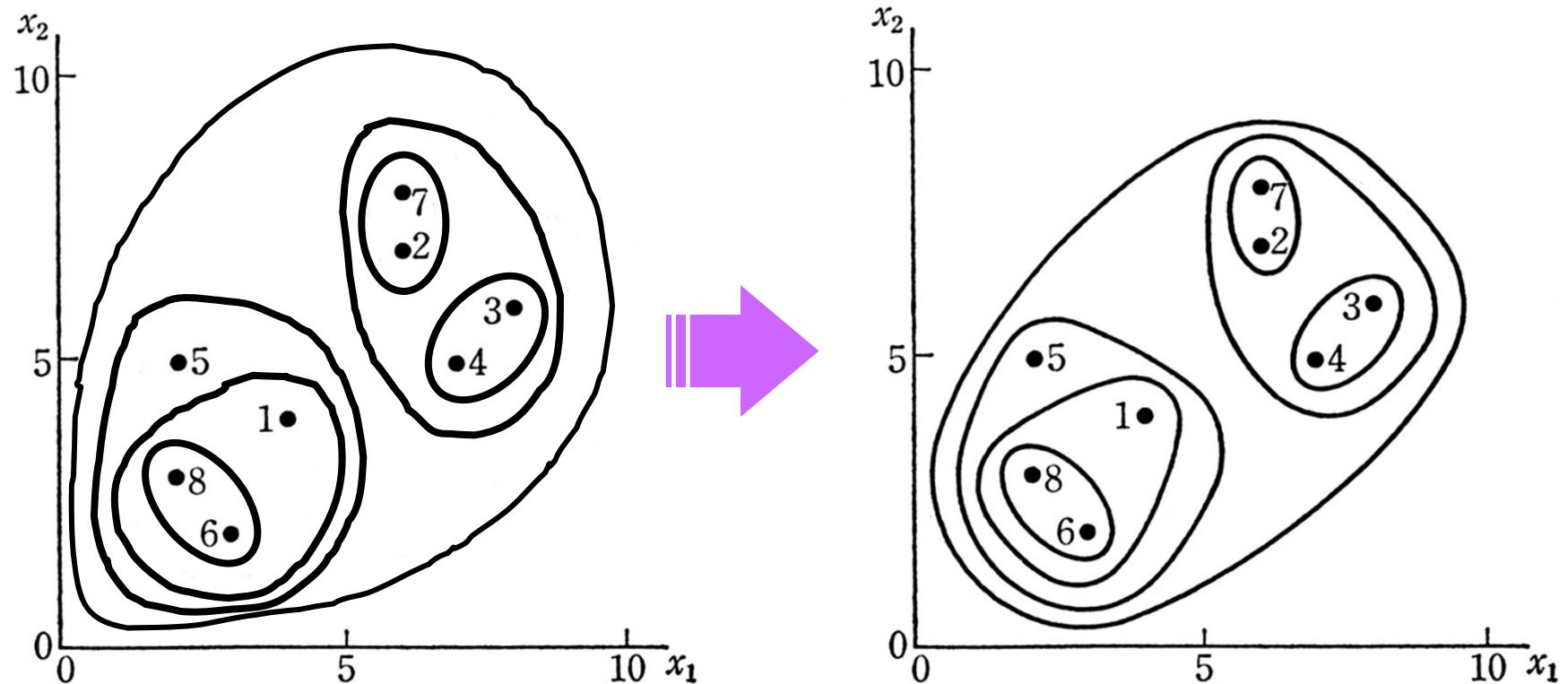
凝集型階層的分類法の簡単な例

- 同じ2変量のトイ・データで模式的に示す.
- 散布図内の各点の間の“非類似度”, “類似度”などを約束する.
- ここでは, 非類似度としてユークリッド距離, 平方ユークリッド距離などを考えればよい.
- 距離が“近い点”(似ているデータ)から順に併合する (agglomeration, lumping, merging)を行なう.
- 2つ以上の点が併合されたときを“クラスター”とする.

(つづき)

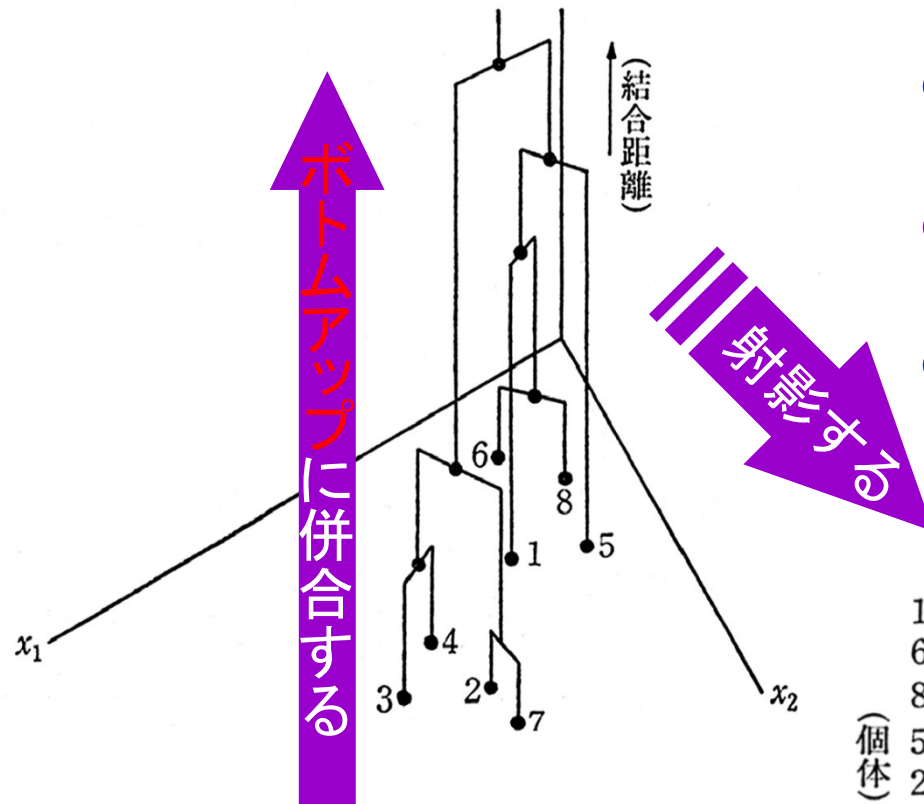
- クラスター間の非類似度, 類似度も約束し(つまりクラスター間の似ている程度), これを使って併合を反復する.
- 似ている(=類似度), 似ていない(=非類似度)の設定(仮定の仕方)によって結果が異なることがある.「どう約束するのか?」という問題がある.
- 一見, 簡単なようだが, 細かい問題がいろいろある.
 - 類似度・非類似度(例: 距離)の選択, 適用可能性
 - 近傍の探し方[近い, 遠い, の約束と探し方]
 - 結合順と標識の付け方(ラベリング)
 - 同値の距離の処理(タイ・ブレイキング), 等々

階層的分類法: クラスター生成の例

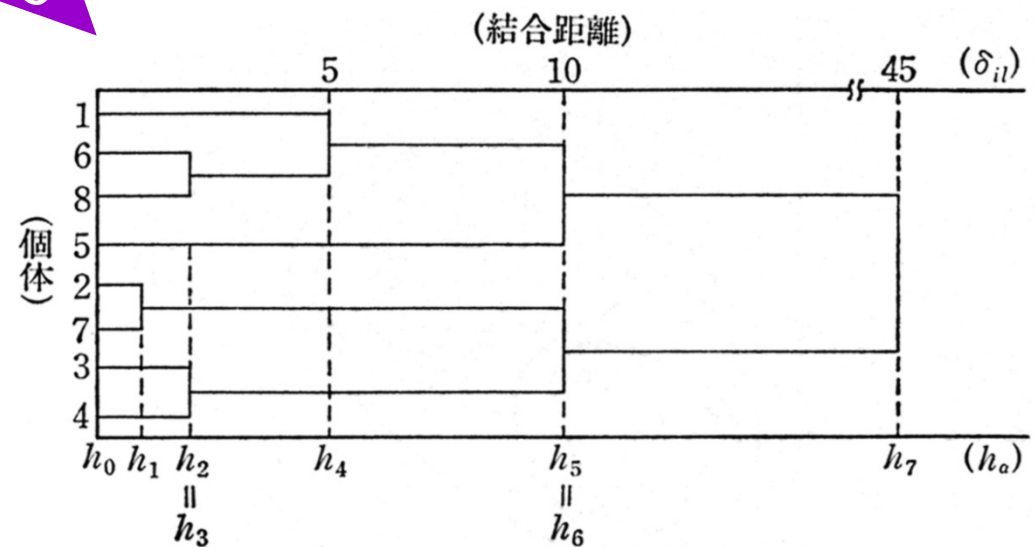


包含関係, つまり階層構造となる
この関係进行操作手順に従って, 模式図としてみると, ...

階層的分類法の考え方(模式図)



- 右の樹形図と始めのデータの散布図と比べてみよう.
- 2変量以上, 多変量ではどう考えるのか.
- 樹形図も1つの“近似表現”である.



階層的分類法の考え方

- 散布図の点の“非類似度”（距離）の大きさに従って、ボトムアップに併合を繰り返す.
- 近い距離にあれば高さは低く、遠い距離になるほど高さが高くなるだろう.
- これを階層的に表すと図のような“階層構造”となる.
- これを“樹形図(デンドログラム)”という.
- 各点を見ると散布図の布置図の実感に合っているように見える(ここは2次元だから、一般に本当か?).
- これを多変量データすると、何か構造があれば、その特徴が樹形図に現れるのではないかと期待がある.

★用いる階層的分類法の違いを知る

- 3種類のトイ・データで比べる(1つは同じ).
- 組み合わせ的手法で表せる階層的分類法の2つ“最短距離法”(single linkage)と“ワード法”を比べる.
 - データA: 大きさが $n=8$ の2変量データ
 - データB: 大きさが $n=15$ の2変量データ
 - データC: 大きさが $n=30$ の3変量データ

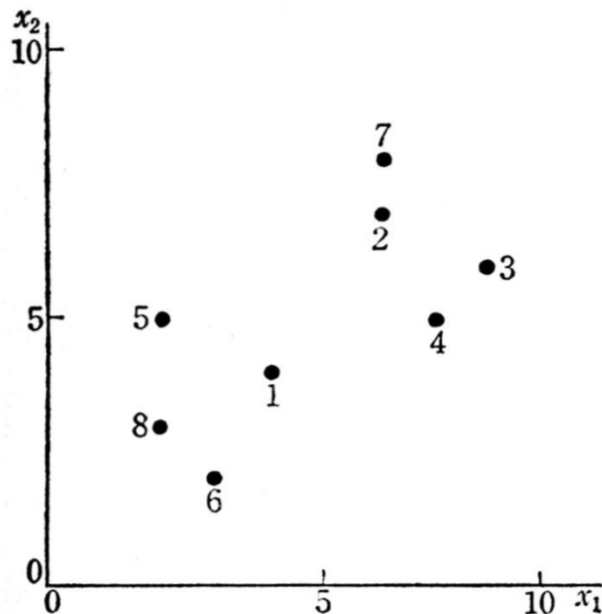
比較的構造が明らかなデータと曖昧なデータ
- 手法により異なる“クラスター化”の特徴を眺める
- つまりクラスターは“生成されるもの”を知る(クラスター化).
- クラスター化法の特徴として, 同じデータセットから出発しても, 1つの解とはならないことがある.

最短距離法のさまざまな別称: single link(単連結法), single linkage, nearest neighbor(最近隣法)

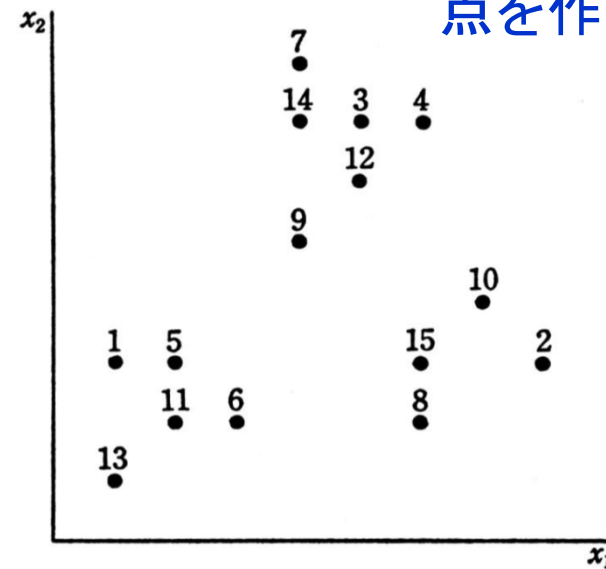
データの様子

- 確認1: どんな分類結果となるか？
- 確認2: クラスタとはなにか(どう作られたか)？
- 確認3: 手法(≡アルゴリズム)の違いは？

注意: 意図的に
“等距離”になる
点を作っている



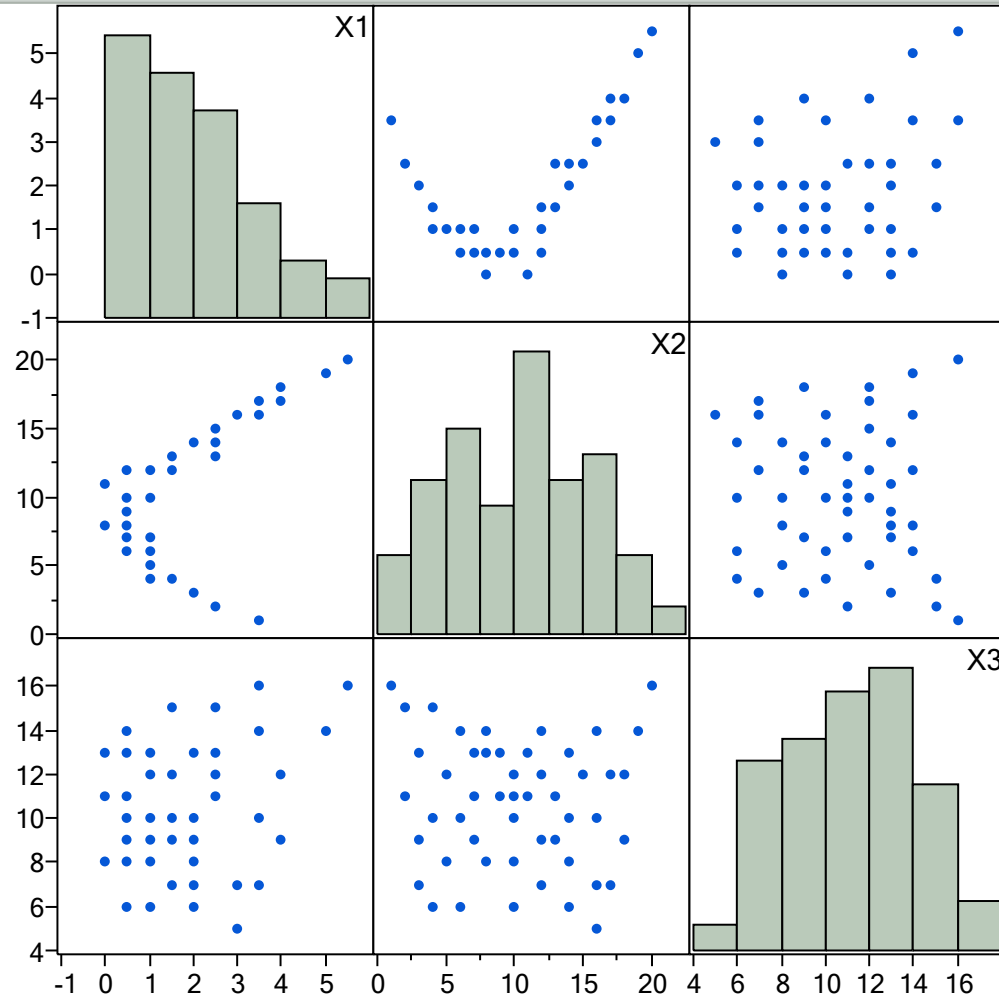
データA



データB

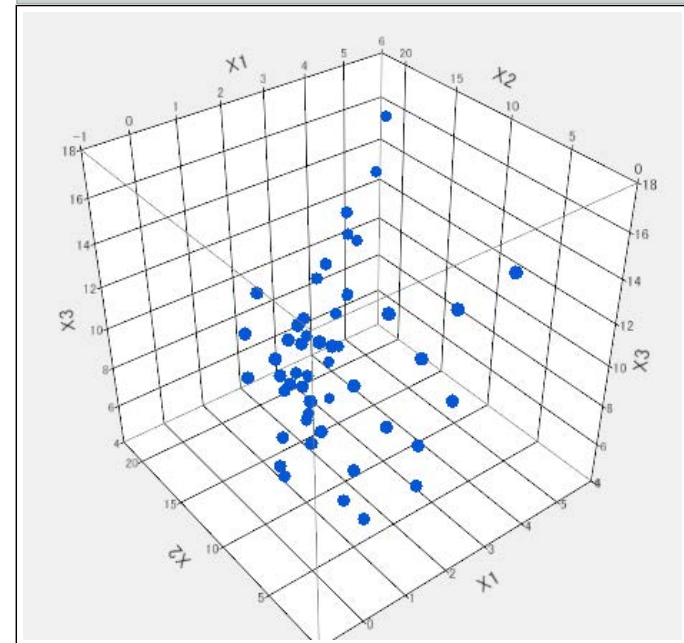
「データC」の散布図行列と3次元散布図

散布図行列



データC

三次元散布図



データ列 X1 X2 X3

確認

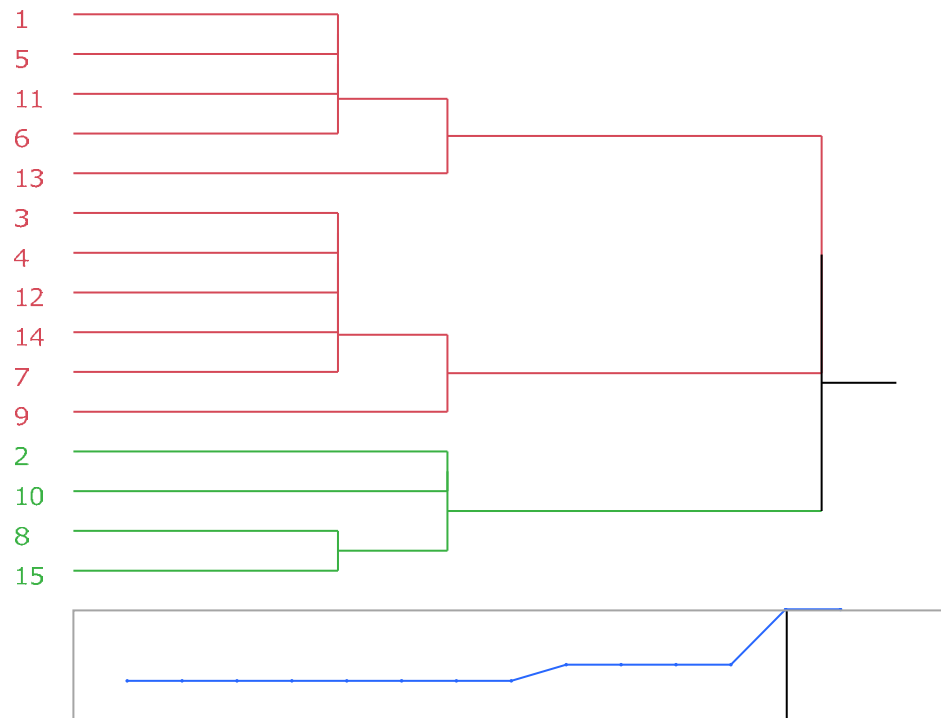
ある構造・形があるが検出できるのか

「データB」デンドログラム:最短距離法

階層型クラスター分析

手法 = 最短距離法

樹形図



連鎖現象・鎖状効果(chaining effect)

階層型クラスター分析

手法 = 最短距離法

クラスター分析の履歴

クラスターの数	距離	結合先	結合者
14	1.000000000	3	4
13	1.000000000	1	5
12	1.000000000	1	11
11	1.000000000	1	6
10	1.000000000	3	12
9	1.000000000	3	14
8	1.000000000	3	7
7	1.000000000	8	15
6	1.414213562	3	9
5	1.414213562	2	10
4	1.414213562	2	8
3	1.414213562	1	13
2	2.828427125	1	3
1	2.828427125	1	2

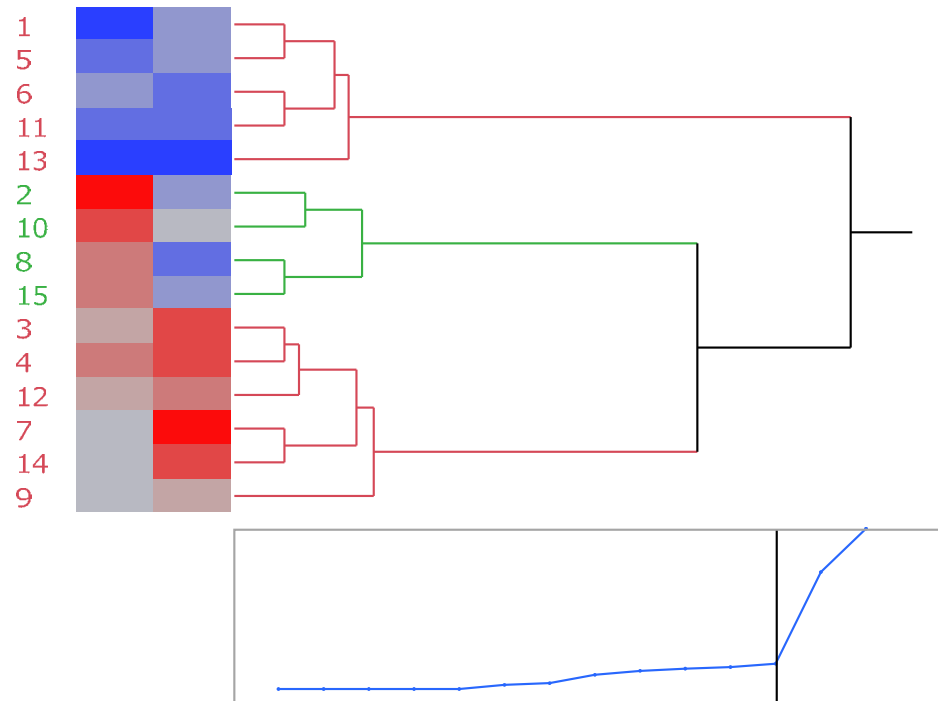
↑↑↑
この出力情報
に注目

「データB」デンドログラム:ウォード法

階層型クラスター分析

手法 = Ward法

樹形図



階層型クラスター分析

手法 = Ward法

クラスター分析の履歴

クラスターの数	距離	結合先	結合者
14	0.707106781	3	4
13	0.707106781	1	5
12	0.707106781	6	11
11	0.707106781	7	14
10	0.707106781	8	15
9	0.912870929	3	12
8	1.000000000	2	10
7	1.414213562	1	6
6	1.612451550	1	13
5	1.722401424	3	7
4	1.802775638	2	8
3	1.966384161	3	9
2	6.533248299	2	3
1	8.696742685	1	2

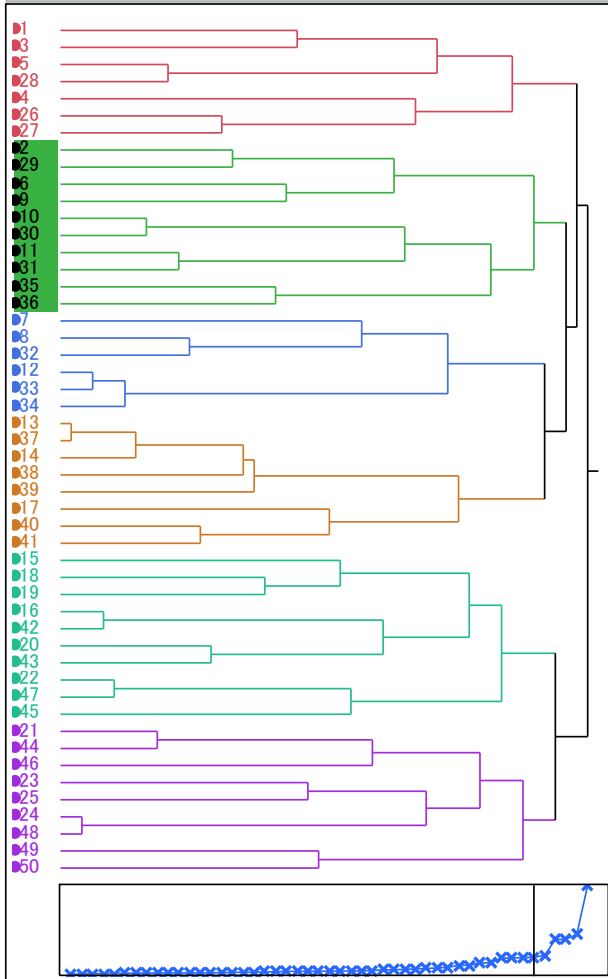
↑ここに注目
前ページと比較

「データC」デンドログラムの比較

階層型クラスター分析

手法 = Ward法

樹形図

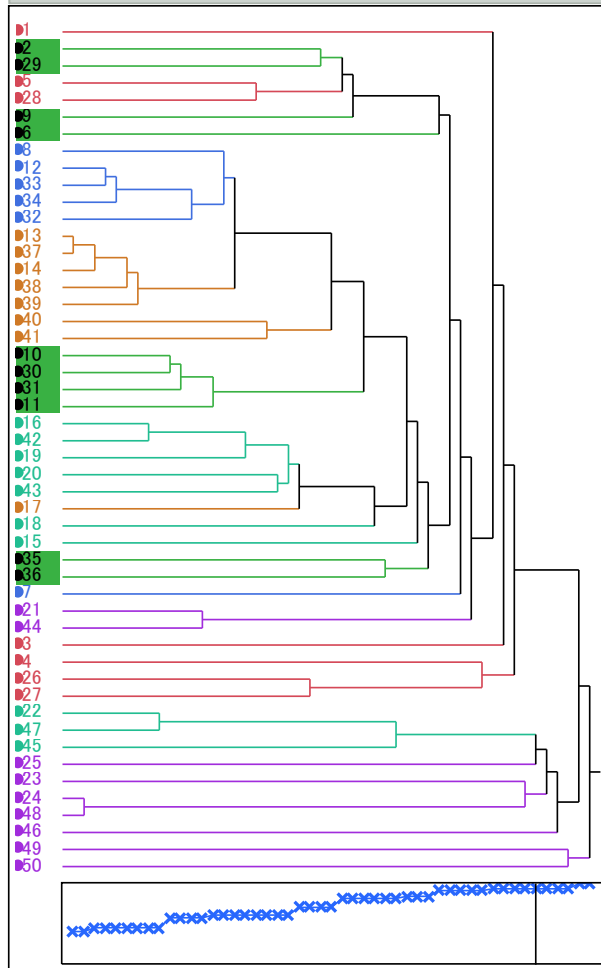


ワード法

階層型クラスター分析

手法 = 最短距離法

樹形図



最短距離法

「データC」デンドログラムの比較

- どちらも「6群」とし、クラスターを色で分けた.
- ウォード法の1つのクラスターを「緑色」で表示.
- 2つの手法のクラスター化過程(結合順とクラスター化情報)とIクラスター構造が異なる.
- 最短距離法にはいわゆる“連鎖現象”がある.
- ウォード法には、見かけの“分離現象”がある. つまり、よく分かれているようにみえる).
- どちらも正しい. つまり、クラスター化基準と併合のアルゴリズムが異なるから、それに合ったクラスターを“生成”した.

遊びで1つデモ(このデータCを対応分析で見ると)

ここのクラスター化の特徴

- (フランス流の)対応分析では, ワード法と k -平均法を併用するクラスター化手順を用いる. これを“ハイブリッド法”(混合方式)と呼んでおく.
- 成分スコアを用いて分類を行う. 2つの指定方法があり, 総変動(全慣性)が異なるので注意する^(†)(後述).
 - 全成分数($K = \min\{m, n\} - 1$)を用いる場合
 - 指定した成分数($K^* < K$)を用いる場合
- ワード法を用いる際に, “相互最近隣の規則”(RNN)を用い, 近い位置にある点(似ている成分スコア)の圧縮化処理を行う.

ワード法と相互最近隣の規則, k -平均法, 用いる成分数
相互最近隣の規則(RNN: reciprocal nearest neighbors rule)
成分スコアの成分数指定:

(†)①JMP「CAスクリプト」は全成分／②WordMinerは上の2種

最近隣 (NN) と相互最近隣 (RNN)

- ある点 (成分スコア) からみて一番近くにある, つまり“最近隣” (NN) にある点を探す.
- 最近隣の基準は, いろいろな分野で利用されてきた.
- その相手の点からみても, ある点がもっとも近いときを“相互最近隣” の関係にあるという.
- こういう関係にある点は, より近い関係にある (よく似ている) とみなせる.
- こうした点を先に集めることで階層化の作業量を低減させることができる. 点の分布に依存する.

.....
NN (nearest neighbor), nearest-neighbor chainもある
RNNはMcQuitty (1966) が考えた.
相互最近隣をmutual nearest neighbor (MNN) ともいう.
.....

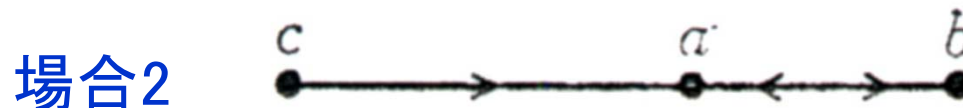
相互最近隣 (RNNまたはMNN)とは

- つまり、点の分布の密集度(稠密度)に依存する.
 - 多くの場合、塊状(房状)の複数のクラスターがあることはあまり期待できない. そもそも分布形状がわからない.
 - 点が密集する機会が多いような場面もある. そのとき計算効率化が期待できる.
 - “相互最近隣”を、模式化する.
-
- 相互最近隣 (R NNまたはMNN)とは、ある2つの点 a, b について、 a が b の最近隣になり、かつ b が a の最近隣でもあるときをいう.
 - 点 a について、その最近隣にある点 b を $b = \text{NN}(a)$ で表す. このときRNNとは、 $b = \text{NN}(a)$ かつ $a = \text{NN}(b)$ のときをいう.
 - 次ページに模式図を付けた.

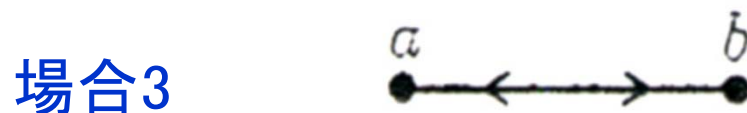
(つづき)



点 a は b の最近隣,
 b はそうでない.



点 a と b は相互最近
隣, c はそうでない.



点 a と b とは厳密な
意味で最近隣.

- 最近隣にある関係を表すグラフをNN-graphといい, えられる連鎖をNN-chainという.
- NN-chainを作って, このグラフ上でRNNを探すアルゴリズムもある.

クラスター化手順の概要

- ここでは次の手順でクラスター化を行う(要点のみ).
 - ① “相互最近隣の規則”を併用するワード法で“予備分類・分割”(preliminary partition)を行う.
 - ② 結合水準やその結合水準の和(クラスター内変動の和に相当)を参考にクラスター数を指定する.
 - ③ 指定したクラスター数を“初期分類”とし, すべての要素(分類対象)の“再分類”(であり細分類)を行う.

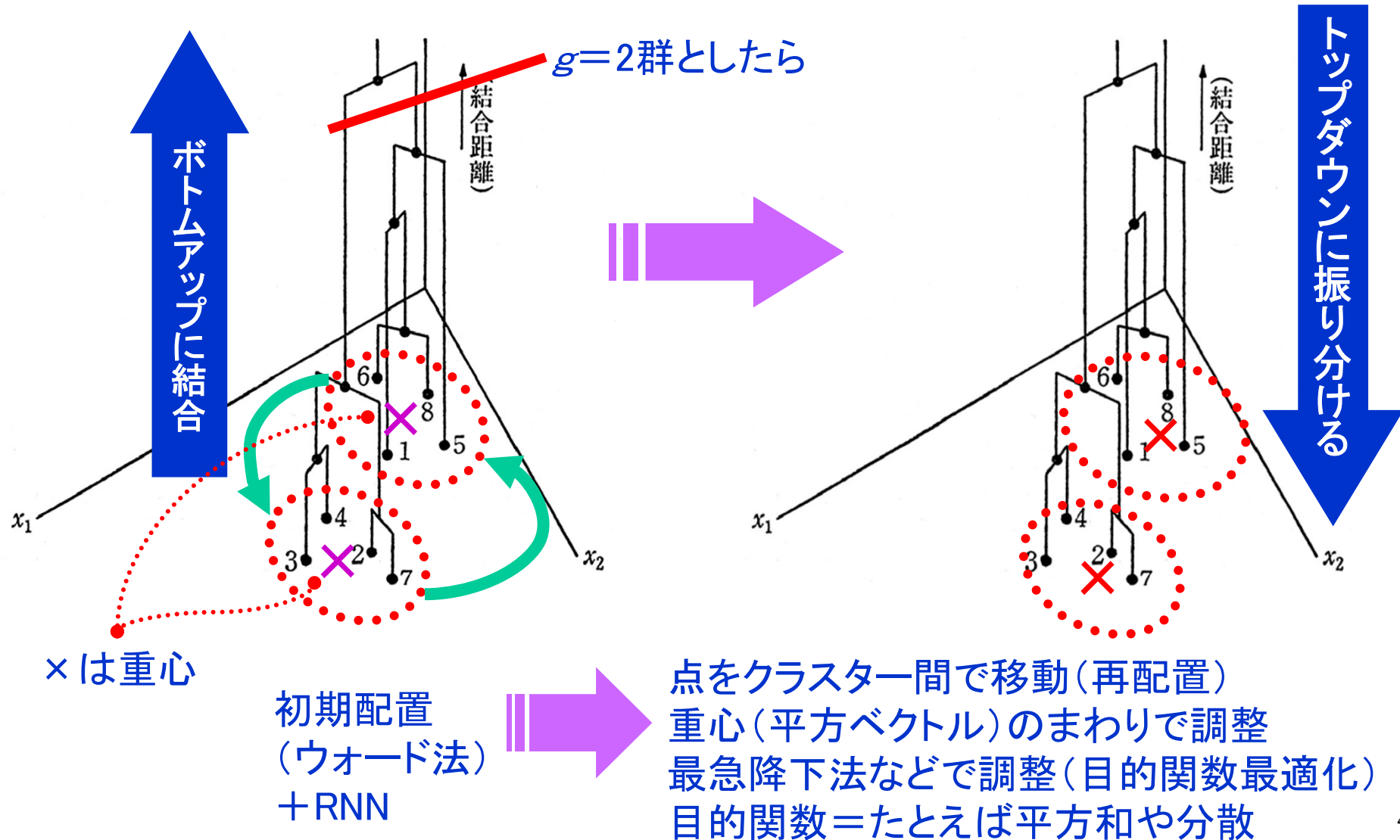
この手順は, WordMinerと対応分析に固有の方法だが, 容易に思い付く方式なので類似方法が他にもある.

(つづき)

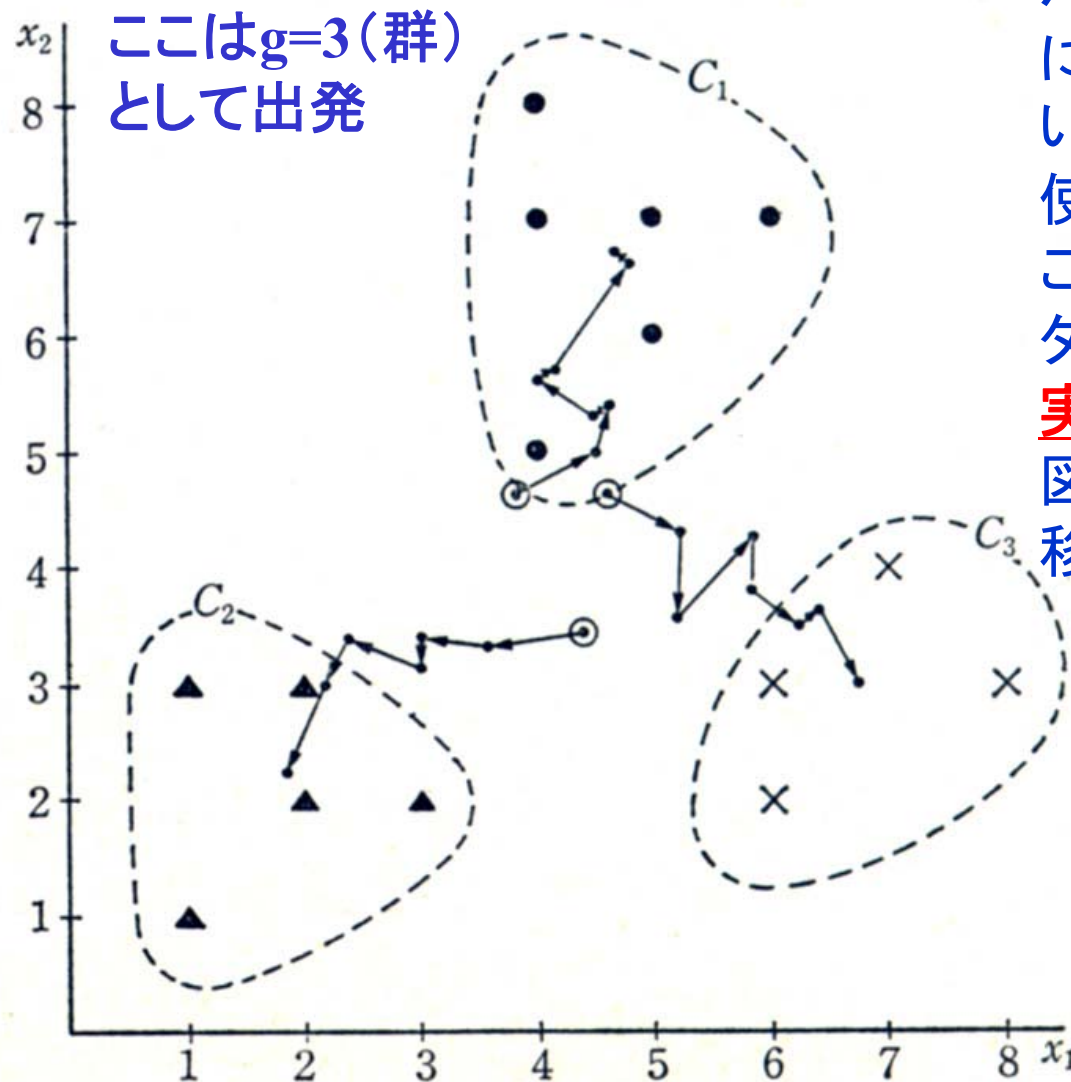
④“初期分類”に対して、つぎに k -平均法で、各クラスターの重心のまわりに“重心移動アルゴリズム”(moving center algorithm)の“再配置法”(reallocation)を適用、なるべくクラスター内変動を小さくなるようクラスター内の点の移動・調整(consolidation, refinement)を行う。

- 成分数($K = \min\{m, n\} - 1$)、ユーザー指定成分数 K^* ($< K$)やクラスター数(g)の指定は、デフォルト値を設ける。
- ここで、 K^* ($< K$)の指定、クラスター数の指定で、統計量に違いがあるので“解釈に注意”する。
- これらは順をおって要点を説明する。

ハイブリッド法のイメージ



k -平均法のみによるクラスター化(イメージ)



ハイブリッド法では, 初期化に「ワード法+RNN」を使い, 細分類に「 k -平均法」を使う.

この例のように“クラスター”(塊)が明確な例は現実にはほとんどない.

図は, 初期配置の重心の移動と収束までの履歴.

(つづき)

- ① クラスタ数(g)を指定する. たとえば, $g=3$ (群)とする.
- ② 初期化する(初期配置を決める). 一般に分類結果は初期化に依存する. 解は一意でない.
ランダム配置, 系統的配置, 恣意的配置などがある.
- ③ クラスタ間の点の移動とクラスタ化基準(たとえば“クラスタ内平方和の和”)の最小化を行う.
- ④ これが小さくなる方へ“山下り法”(急降下法)などで最適化する.
- ⑤ 一意の最適解とならず“局所的最適”となる.
分割の組合せのすべてを網羅的に調べられない.

最適解を求めることは, 現在のスーパーコンピュータでも, たぶん難しい. クラスタリングにはこの種の課題が多い.

ワード基準によるクラスター化(概要)

- ここで, “2元データ表”から出発するとする.
- この対応分析法の結果でえた(行または列の)“成分スコア”を用いたクラスター化を考える.
- これは“カイ二乗距離”を使うことに同じである.
- 成分スコアについては(平方)ユークリッド距離が適用できるので“ワード基準”によるクラスター化を行う.
- また, クラスター化時には, 得られた“全成分”(K)を用いることを基本とする^(†).

JMPの標準機能には「高速ワード法」がある. これは(たぶん)RNNと類似の手順を階層的分類で用いている.

(†)これを変えた場合も重要. うしろで説明する.

(つづき)

- カイ二乗統計量の分解を用いるクラスター化とワード基準によるそれとは、同じことを言い換えただけ.
- 2元データ表のセル内の度数が非常に疎で、しかも寸法が大きいと、クラスター化計算にやや時間を要する.
- 必要に応じて“デンドログラム”(樹形図)を出力表示する. データ表の寸法が大きいと視覚化の限界がある.
- WordMinerは規模の大きいデータセットを想定しているのでデンドログラムは出力しない(クラスター化履歴情報は出力する).
- これらを取り混ぜて、レストランデータで要点を示す.

対応分析法とクラスター化法

- 再び「レストラン評価」のトイ・データを例として調べる.
- 原データ表の「(回答者) × (2項目)」の場合を確認.
- たとえば, 行側の分類として「レストランの名前を自由記述」で得た, と考える.
- 「評価基準」の側(列側)からみても同じこと.
- この2項目のクロス表を用意する. これが所与の“2元データ表”となる.
- ここで“レストランの分類”は, 同時に“回答者”の分類も暗黙に想定していること.
- また列側の“評価項目”の分類もあること.
- 以下「★」印スライドは, 再確認かつ復習.

★データのイメージ(X表)

項目 回答者	I (レストラン)	J (評価基準)
1	バッハ	味
2	ムガール	量
3	さとみ	量
4	ラ・マレ	工夫・サービス
5	きくみ	味
⋮	⋮	⋮
⋮	⋮	⋮
N	いりふね	量
$N=1,284$ (回答者数)		

なんども繰り返すが、これは“質的データ”である。
 回答者が、こういう自由記述を行った、と読み替えてみる
 (実際、こうしても矛盾はない)。

★再確認: 2元クロス表 (F表) $\mathbf{F} = (f_{ij})_{10 \times 3}$

$m = 10, n = 3 \Rightarrow I = \{1, 2, \dots, 10\}, J = \{1, 2, \dots, 3\}$

項目I \ 項目J	工夫・サービス	味	量	行和
いりふね	98	25	32	155
かりや	105	35	38	178
きくみ	35	0	67	110
さとみ	4		7	95
クラーク	3		54	102
コルシカ	3		13	122
バッハ	48	76	18	142
ムガール	49	44	16	109
ラ・マレ	49	82	15	146
ロゴスキー	48	35	42	125
列和	540	442	302	1,284 (=N)

これを行のレストランが列の評価項目の空間内に布
置する多次元データとみ
ている(逆も同じ).

「」を覚えておく(質量に關係)

★対応分析法の性質

- すでに確かめたつぎの2つの性質を再確認する.
- 2元データ表の総変動(全慣性)であり, “全情報”である.

[性質1]

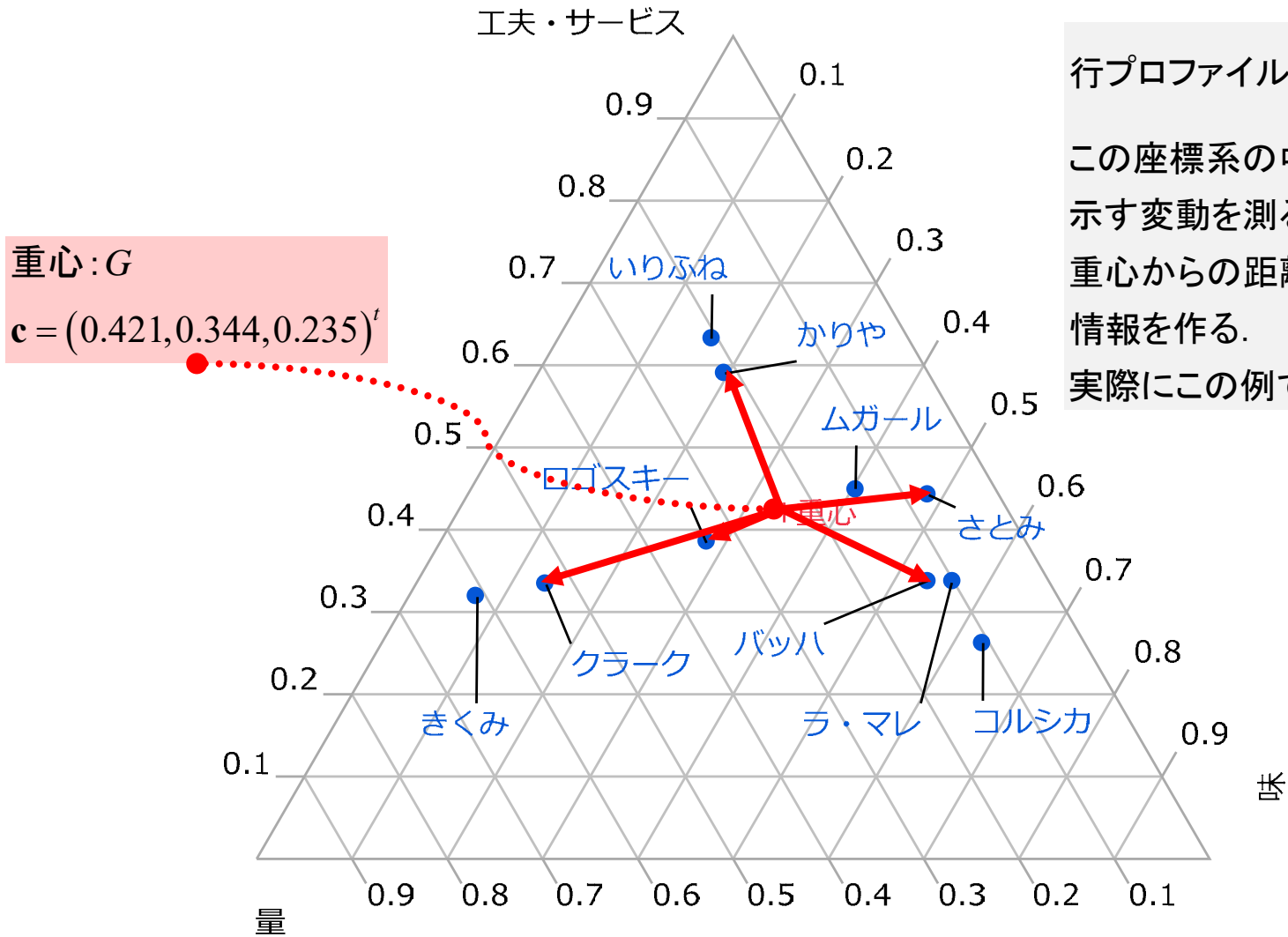
[固有値の和] = [カイ二乗統計量] ÷ [クロス表の総度数]

$$\phi^2 = \sum_{k=1}^K \lambda_k = \frac{\chi_p^2}{N} \quad (\text{ここで, } K = \min\{m, n\} - 1)$$

[レストラン評価データの確認]

$$\sum_{k=1}^K \lambda_k = \lambda_1 + \lambda_2 = 0.2577 \quad \Leftrightarrow \quad \frac{\chi_p^2}{N} = \frac{330.860}{1284} = 0.257679 \dots \doteq 0.2577$$

★「行プロフィール」プロットを観察



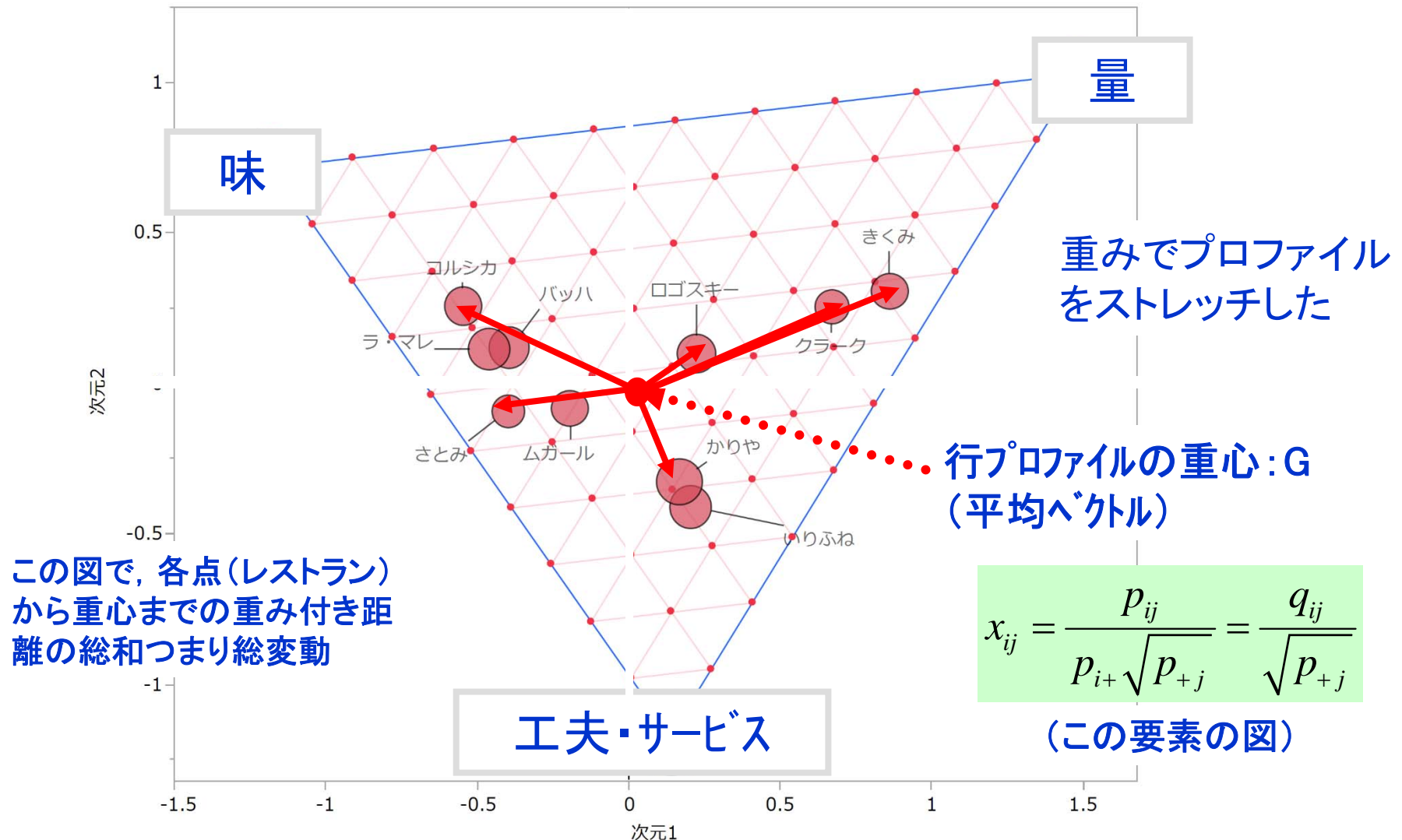
行プロファイル: $q_{ij} = \frac{p_{ij}}{p_{i+}}$ をプロット

この座標系の中で各点(レストラン)が示す変動を測る.

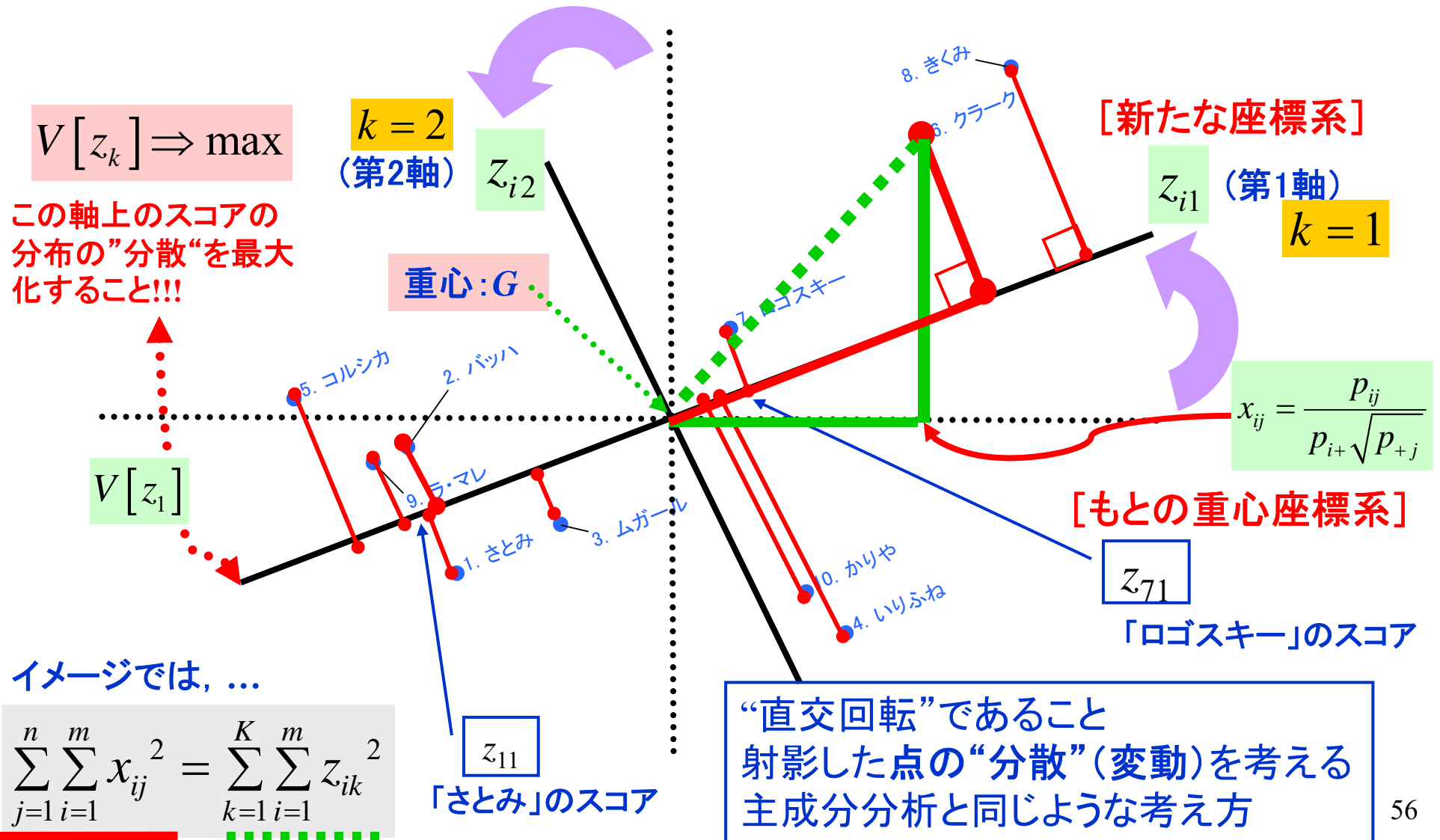
重心からの距離, つまり分散相当の
情報を作る.

実際にこの例でそれを作る(算出)する.

★ストレッチ・プロファイルで観察



★重心座標系から (z_1, z_2) 系へ変換(射影)



★対応分析法の性質

[性質2]

クロス表の第 $i \in I$ あるいは第 $j \in J$ について 以下がなりたつ.

$$\text{総変動(全慣性)}: \phi^2 = \frac{\chi_p^2}{N}$$

$$= \sum_{i=1}^m (\text{クロス表の第}i\text{行の質量}) \times \left[\begin{array}{l} \text{第}i\text{行プロフィールと行の} \\ \text{平均プロフィール(重心)との}\chi^2\text{ 距離} \end{array} \right]$$

$$= \sum_{j=1}^n (\text{クロス表の第}j\text{行の質量}) \times \left[\begin{array}{l} \text{第}j\text{行プロフィールと} \\ \text{列の平均プロフィール(重心)との}\chi^2\text{ 距離} \end{array} \right]$$

- あとで「選択肢(i または j)を“クラスター”と読み替えて用いる.
- 前のトーク内容を確認(「第2回配布」にある).

★数値例:「行側」から全慣性を求める

- たとえば「さとみ」の行プロフィールについて調べる.
- すべてのレストラン(項目*i*)について重み付きの重心との間のカイ二乗距離を求め和を作る. 表にまとめる(次ページ).
- 図で確認する. おおまかに言えば赤い矢印の二乗和となる.

$$\begin{aligned}
 & \left[\begin{array}{l} \text{「さとみ」の質量} \\ \times \text{「さとみ」プロフィールから行重心までのカイ二乗距離} \end{array} \right] \\
 &= \underbrace{0.074}_{\text{質量}} \times \left\{ \underbrace{\frac{(0.442 - 0.421)^2}{0.421}}_{\substack{\text{エ夫・サービス} \\ \text{ここがカイ二乗距離の部分}}} + \underbrace{\frac{(0.484 - 0.344)^2}{0.344}}_{\text{味}} + \underbrace{\frac{(0.074 - 0.235)^2}{0.235}}_{\text{量}} \right\} = 0.01246 \quad (\text{▲})
 \end{aligned}$$

★すべてのレストラン($i \in I$)について確認

表 8 カイ二乗距離の算出ほか(桁数を増やしてリチェック:7 桁で確認)

レストラン名 (i)	カイ二乗距離の要素			構成要素数 構成比	距離	$\frac{\chi_p^2}{N} = \sum_{\alpha} \lambda_{\alpha}$ の確認
	工夫・ サービス	味	量	①列の平均プロフ ファイル(行の質量) (p_{i+})	②平方カイ二乗距離	③=①×②
いりふね	0.1066041	0.0972176	0.0035027	0.1207165	0.2073243	0.0250275
かりや	0.0681846	0.0633187	0.0020025	0.1386293	0.1335058	0.0185078
きくみ	0.0249137	0.2141904	0.5943787	0.0856698	0.8334828	0.0714043
さとみ	0.0011031	0.0569077	0.1108962	0.0739875	0.1689070	0.0124970
クラーク	0.0181054	0.1243993	0.3679900	0.0794393	0.5104948	0.0405533
コルシカ	0.0595549	0.2390521	0.0703164	0.0950156	0.3689235	0.0350535
バッハ	0.0162076	0.1059357	0.0499616	0.1105919	0.1721049	0.0190334
ムガール	0.0019913	0.0102716	0.0332267	0.0848910	0.0454897	0.0038617
ラ・マレ	0.0171636	0.1372508	0.0746459	0.1137072	0.2290603	0.0260458
ロゴスキー	0.0031783	0.0119870	0.0431974	0.0973520	0.0583627	0.0056817
						0.2576660 (固有値の和)

(▲)この「さとみ」を調べた。

↑
(表 5) 参照

(★)個々のレストランの重心からの距離(つまり“変動”)

表の③欄の各レストランの変動, 総変動の関係

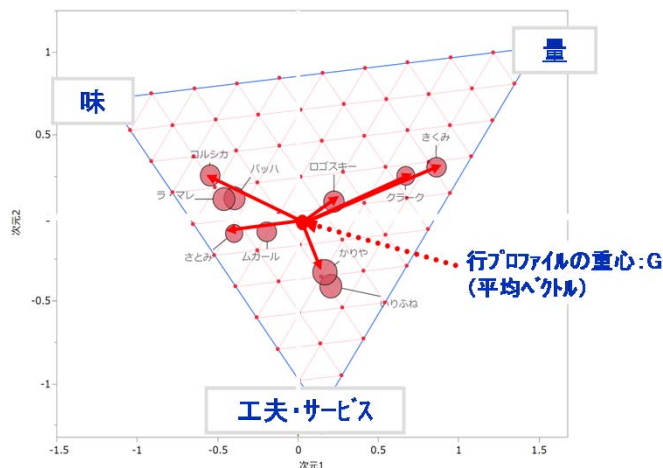
$$\sum_{i=1}^{10} \left[\begin{array}{l} \text{(第}i\text{番目の質量)} \\ \times \text{(第}i\text{番プロフィールから重心までのカイニ乗距離)} \end{array} \right]$$

$$\underbrace{0.02501}_{\text{いりふね}} + \underbrace{0.01818}_{\text{かりや}} + \underbrace{0.07172}_{\text{きくみ}} + \cdots + \underbrace{0.02616}_{\text{ラ・マレ}} + \underbrace{0.00568}_{\text{ロゴスキー}} = 0.2576$$



(★)

$$\phi^2 = \frac{\chi_p^2}{N} = 0.2577 \text{ (総変動であり全慣性)}$$



- この重心座標系(であり三角座標系)の中での点(ストレッチ・プロフィール)の変動を測っている.
- カイニ乗距離を使うのでこうなる.

表の③の欄の意味

- 項目 I の各選択肢(レストラン)の各変動の程度を合わせると“総変動”(全慣性)となる.
- “カイ二乗距離”を使っているのでこうなる.
- ストレッチ・プロフィールの図の赤い線(矢印の二乗)の総和となる.
- ここでは2次元平面で視認できるような例とした.
- しかし, 通常は“多次元”空間である!!!
- その多次元空間内の総変動を, 点(ストレッチ・プロフィール)が, よく分かれる. つまり分散が大きい向きを探しながら, 射影する.

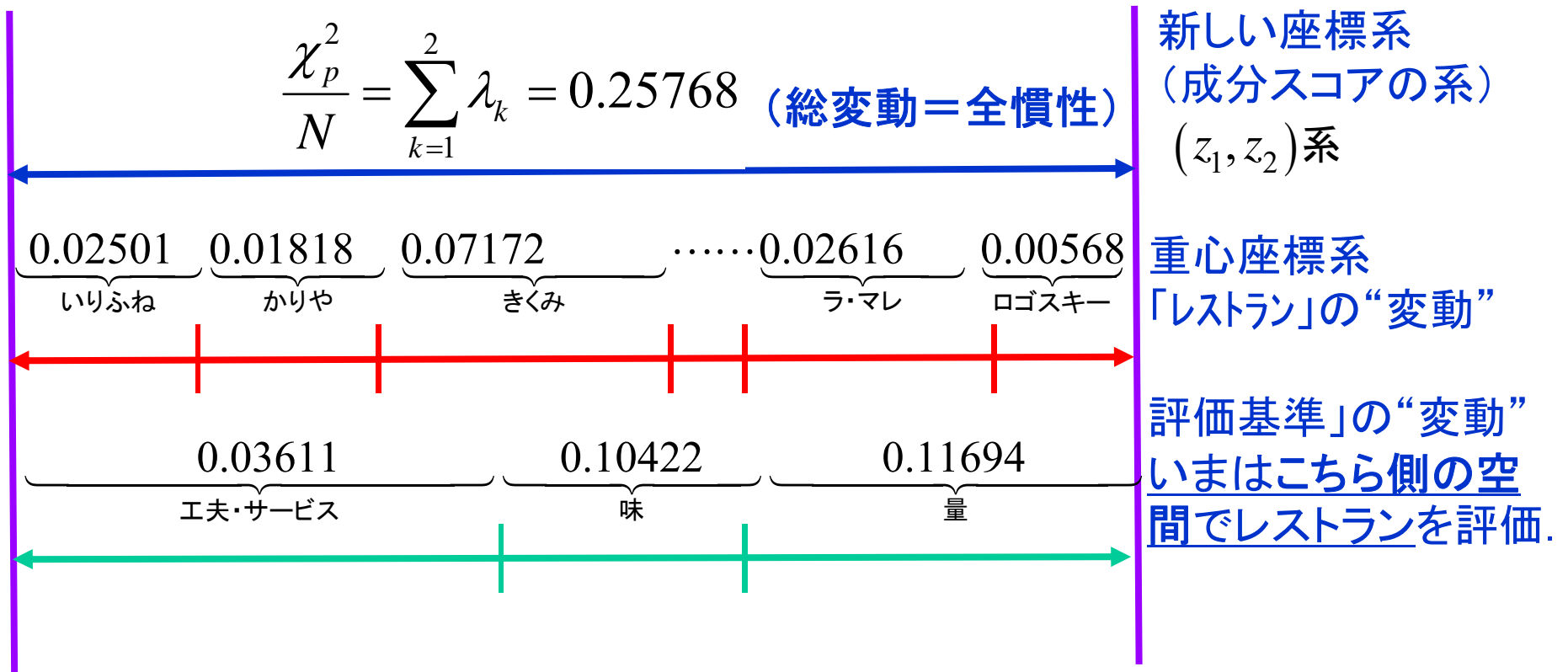
(つづき)

- それを“成分スコアの空間”に転写(射影)し直し, そのあらたな空間内での変動(固有値＝成分スコアの分散)に“読み替える”.
- (分散で測れる)構造があれば, 変動の大きい方から並ぶ(ことが期待される).
- その成分スコアの変動は, 分散(固有値)の大きい順に並ぶようにした!(図の回転と射影の意味)!!!

(つづき)

- ここでは、行側 ($m=10$ のレストラン) が、列側 ($n=3$, 実際は2) の次元数の空間内にあるとしたが、列側が行側の空間に分布すると考えても同じ.
- こうなるのも ストレッチ・プロファイル や カイ二乗距離 を考えることでなり立つ.

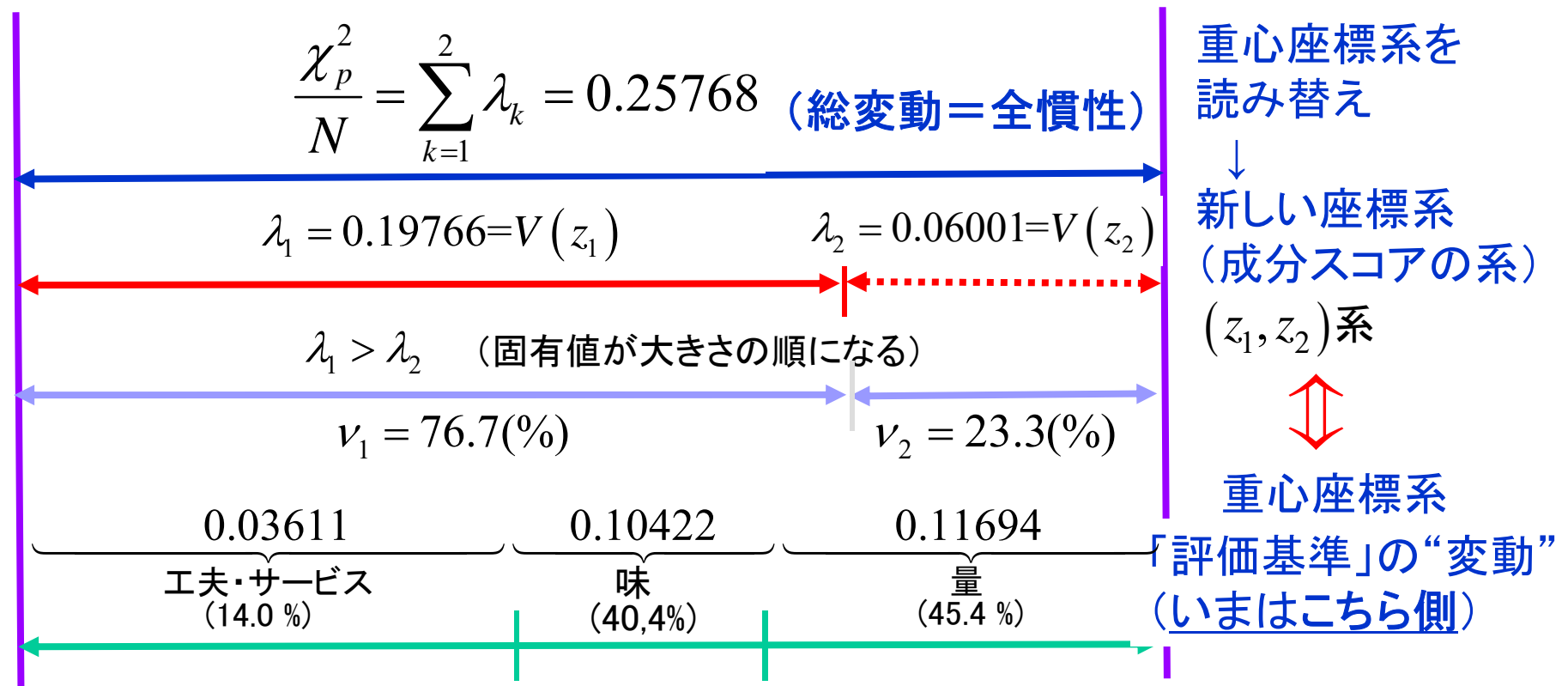
★総変動(全慣性)と行・列の変動の関係



つまり、何度もくどいが…

- ここで求めた $[(\text{質量}) \times (\text{重心までのカイ二乗距離})]$,
つまり個々の選択肢の変動を引用し図に描いた.
- 行(レストラン)と列(評価基準)の両方にある.
- 全体の和は”全慣性“に一致する.
- 総変動という情報を行のレストランあるいは列の評価基準の各選択肢に対応させて分解した.
- いま, レストランが布置する重心座標系の(列側の)評価基準の空間内の変動が示す大きさを, あらたな成分スコアの空間内の変動に読み替える.

総変動と列側（評価基準）の変動の関係



再度確認: レストラン($i \in I$)について確認

表 8 カイ二乗距離の算出ほか(桁数を増やしてリチェック:7 桁で確認)

レストラン名 (i)	カイ二乗距離の要素			構成要素数 構成比	距離	$\frac{\chi_p^2}{N} = \sum_{\alpha} \lambda_{\alpha}$ の確認
	工夫・ サービス	味	量	①列の平均プロファイル (行の質量) (p_{i+})	②平方カイ二乗距離	③=①×②
いりふね	0.1066041	0.0972176	0.0035027	0.1207165	0.2073243	0.0250275
かりや	0.0681846	0.0633187	0.0020025	0.1386293	0.1335058	0.0185078
きくみ	0.0249137	0.2141904	0.5943787	0.0856698	0.8334828	0.0714043
さとみ	0.0011031	0.0569077	0.1108962	0.0739875	0.1689070	0.0124970
クラーク	0.0181054	0.1243993	0.3679900	0.0794393	0.5104948	0.0405533
コルシカ	0.0595549	0.2390521	0.0703164	0.0950156	0.3689235	0.0350535
バツハ	0.0162076	0.1059357	0.0499616	0.1105919	0.1721049	0.0190334
ムガール	0.0019913	0.0102716	0.0332267	0.0848910	0.0454897	0.0038617
ラ・マレ	0.0171636	0.1372508	0.0746459	0.1137072	0.2290603	0.0260458
ロゴスキー	0.0031783	0.0119870	0.0431974	0.0973520	0.0583627	0.0056817
						0.2576660 (固有値の和)

選択肢(j)についても同様の関係

↑
(表 5) 参照

個々のレストランの重心からの距離(つまり変動)

ここでカイ二乗距離とその性質を再確認

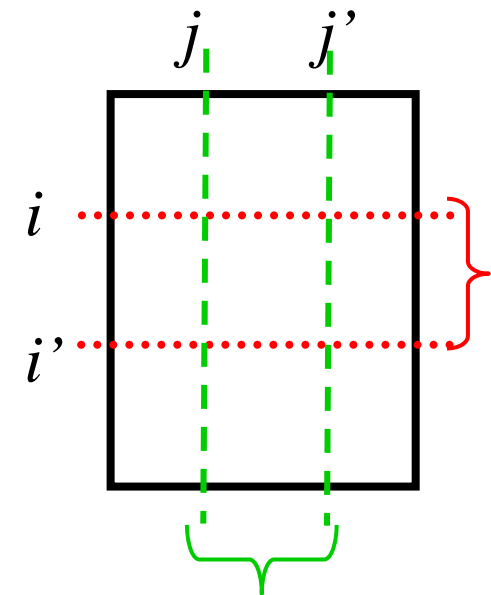
- クラスタ化に必要な距離を考える.
- 総変動(慣性)を考えるとき, 別の見方としてプロファイル間の“カイ二乗距離”がある.
- すでに単純な(平方)ユークリッド距離とどう異なるかを調べた.
- 行のストレッチ・プロファイル間, 列のストレッチ・プロファイル間, それぞれのカイ二乗距離は以下となる.

ここでカイ二乗距離とその性質を再確認

- 行側, 列側の選択肢について, 距離を用意する.
- ここで, カイ二乗距離としないと成分スコアがユークリッド距離となる関係が得られない.

$$\text{.....} \quad d_B^2(i, i') = \sum_{j=1}^n \frac{1}{p_{+j}} (q_{ij} - q_{i'j})^2 = \sum_{j=1}^n \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2$$

$$\text{---} \quad d_B^2(j, j') = \sum_{i=1}^m \frac{1}{p_{i+}} (q_{ij}^* - q_{ij'}^*)^2 = \sum_{i=1}^m \frac{1}{p_{i+}} \left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}} \right)^2$$



カイ二乗距離とユークリッド距離の関係

- ある2元データ表の対応分析で得た“行成分スコア”，“列成分スコア”を用いたクラスター化を考える.
- もとの2元データ表のプロフィール間のカイ二乗距離は，成分スコアのユークリッド距離に“等しい”という重要な性質がある. [テキスト, 第Ⅱ部, 41～42ページに証明あり]

[性質3]

[成分スコア間の平方ユークリッド距離]

= [元のクロス表のプロフィール間の平方カイ二乗距離]

$$d_E^2(i, i') = \sum_{k=1}^K (z_{ik} - z_{i'k})^2 \Leftrightarrow d_B^2(i, i') = \sum_{j=1}^n \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2$$

$$d_E^2(j, j') = \sum_{k=1}^K (z_{kj}^* - z_{kj'}^*)^2 \Leftrightarrow d_B^2(j, j') = \sum_{i=1}^m \frac{1}{p_{i+}} \left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}} \right)^2$$

この目標: レストラン($i \in I$)の分類

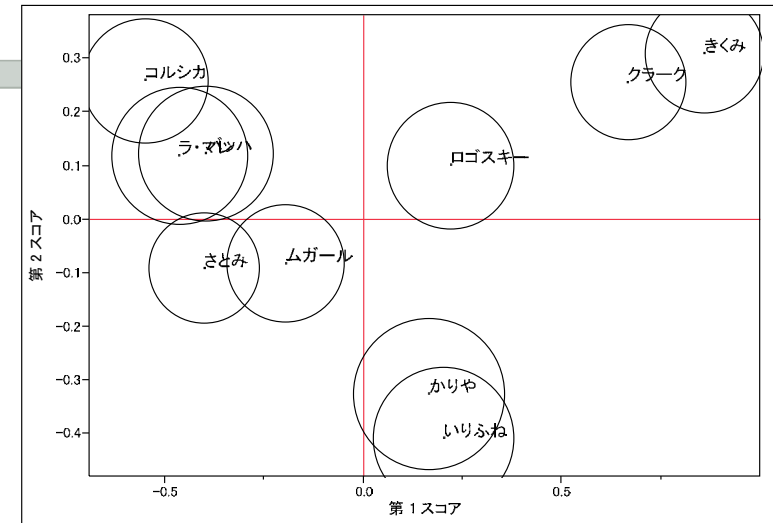
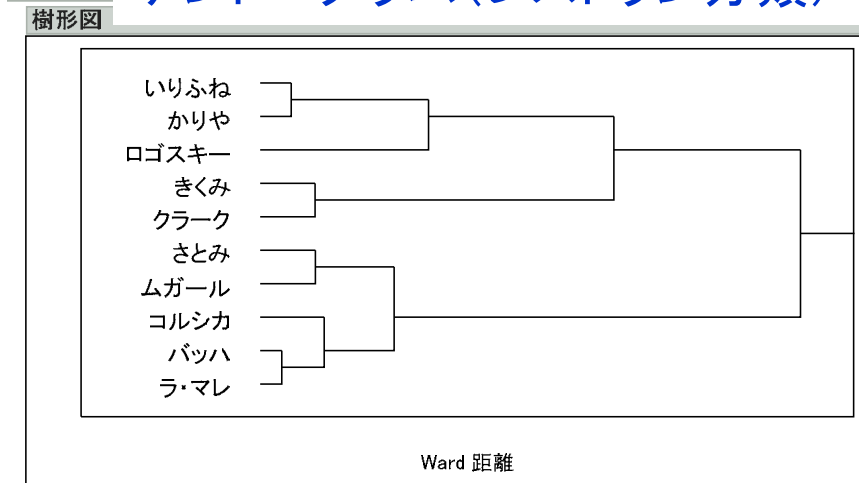
- ここではデータ表の行側, レストランの分類を考える. 列側(評価基準)の分類も同じように行える(対応分析法の特徴).
- 平方カイ二乗距離を用いたクラスター化と, 平方ユークリッド距離を用いたクラスター化は同じである.
- まず, 成分スコアを用いた“ワード基準”によるクラスター化でえられる結果を観察する.
- つぎに, クラスター化過程でえられる“階層の結合水準”と“カイ二乗統計量”の関係を調べる.

(つづき)

- クラスタ化時に用いる“成分数” ($K = \min\{m, n\} - 1$) を, $K^* < K$ と変えたときの性質を調べる.
- つまり, 最大成分数までを用いないときがある.
- 理由は, 多次元データの合成変数であるはじめのほうの少数成分で観察したい(次元縮約したい).
- そのときの総変動と成分の分散(固有値)の関係を知ること.
- 統計量のうち, “総変動”, “クラスター間変動”, “クラスター内変動”, “階層の結合水準”の関係を調べる.
- 実際にソフト(WordMiner, JMP)の出力結果で確認.

レストランの分類結果 (JMPスクリプト出力)

行スコア デンドログラム (レストラン分類)



成分スコア布置図

クラスター分析履歴							
クラスター数	Ward 距離	Ward 距離の2乗	併合先の群	併合された群	併合先の周辺割合(%)	併合された周辺割合(%)	
9	0.01588	0.00025	バツハ	ラ・マレ	11.1	11.1	11.4
8	0.02287	0.00052	いりふね	かりや	12.1	12.1	13.9
7	0.04042	0.00163	きくみ	クラーク	8.6	8.6	7.9
6	0.0406	0.00165	さとみ	ムガール	7.4	7.4	8.5
5	0.04712	0.00222	コルシカ	バツハ	9.5	9.5	22.4
4	0.09863	0.00973	さとみ	コルシカ	15.9	15.9	31.9
3	0.12401	0.01538	いりふね	ログスキー	25.9	25.9	9.7
2	0.26054	0.06788	いりふね	きくみ	35.7	35.7	16.5
1	0.39801	0.15842	いりふね	さとみ	52.2	52.2	47.8

Ward 距離2乗の合計 0.257678804157674

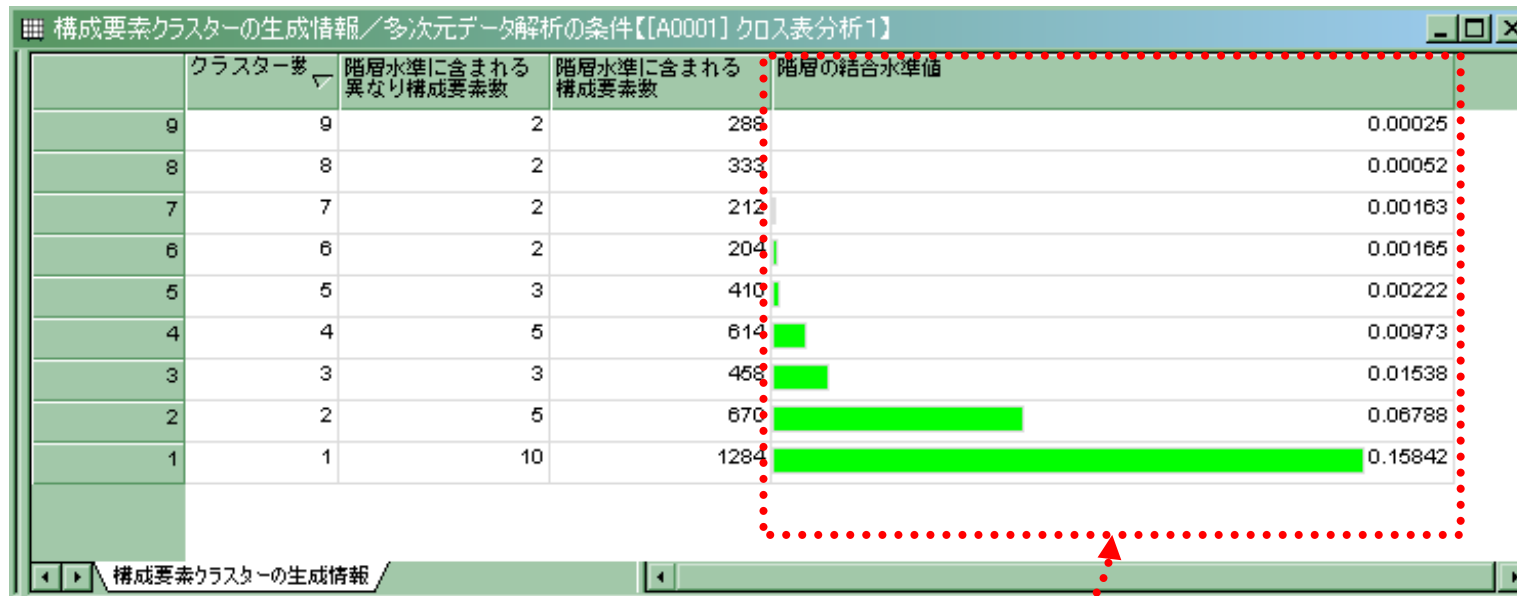
Ward距離2乗の合計=0.25767...=総変動(全慣性)
=階層の結合水準の総和(次ページ)

この結合距離は何か？
(ワード距離の2乗)
次々ページの $h(r)$

73

レストランの分類結果 (WordMiner出力)

- WordMinerでは比較的規模の大きな2元データ表を扱うので、デンドログラムは出力しない。
- クラスタ結合時の“階層の結合水準”と各クラスター内の構成要素(メンバーシップ)を一覧表示する。



この階層の結合水準(前ページのワード距離の2乗, 次ページの $h(r)$)

表に再編集する

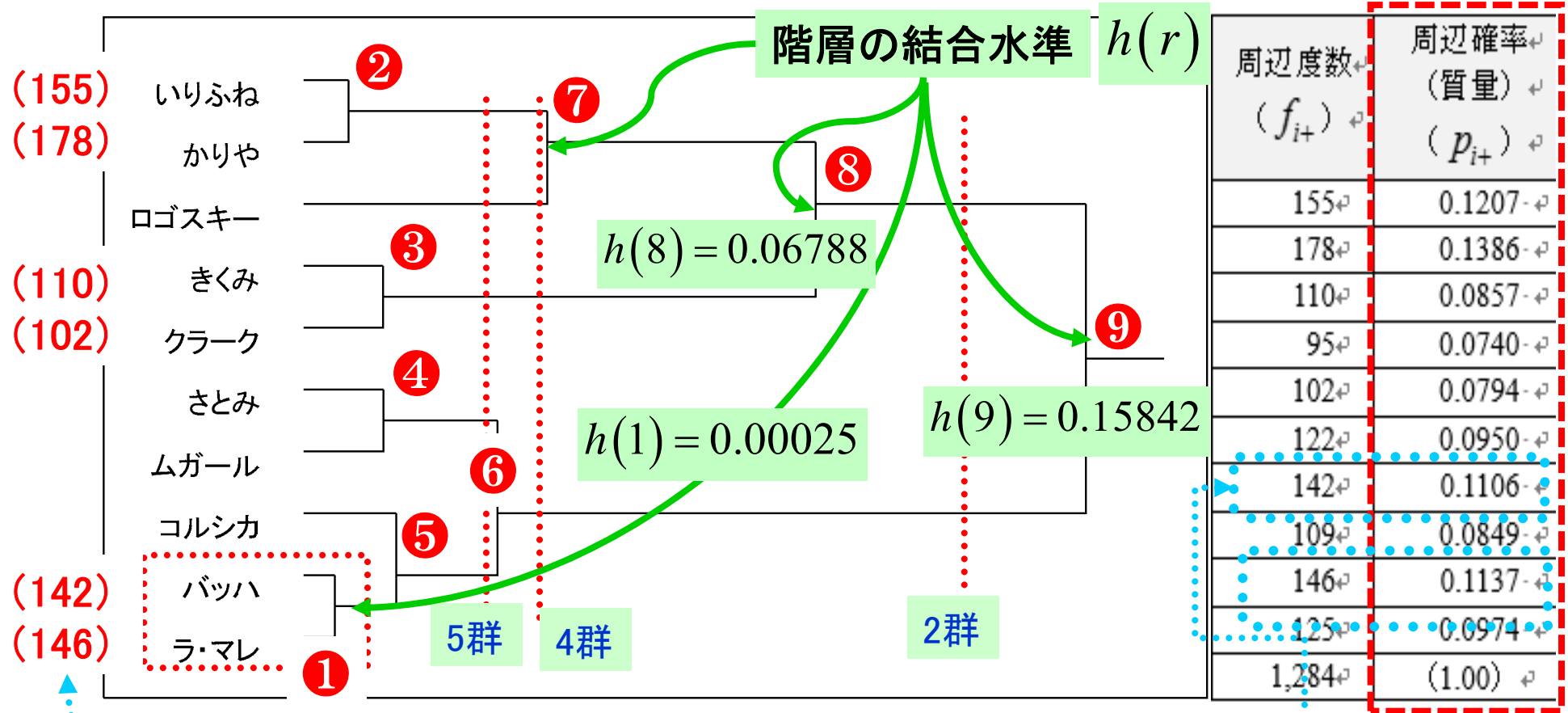
結合の ステップ (r)	クラスター数 (g)	階層水準に 含まれる レストラン数 (t)	階層水準に含まれる 回答者数 (n)	併合先の群 (C_p)	併合された群 (C_q)	階層の 結合水準値 $h(r)$	階層の結合水準値 の累積和 $\sum_r h(r)$
①	9	2	288	パッハ	ラ・マレ	0.0003	0.0003
②	8	2	333	いりふね	かりや	0.0005	0.0008
③	7	2	212	きくみ	クラーケ	0.0016	0.0024
④	6	2	204	さどみ	ムガール	0.0017	0.0041
⑤	5	3	410	ゴルシカ	パッハ	0.0022	0.0063
⑥	4	5	614	さどみ	ゴルシカ	0.0097	0.0160
⑦	3	3	458	いりふね	ロゴスキー	0.0154	0.0314
⑧	2	5	670	いりふね	きくみ	0.0679	0.0993
⑨	1	10	1,284	いりふね	さどみ	0.1584	0.2577
(†) クラスター内に入ったレストラン数 (最後のセルが 10 クラスター) (‡) もとのデータ表からみた, つまり回答者が選んだレストラン数による計数 (最後のセルが 1,284 名の回答者に相当), 表 3 を参照.						0.2577 [結合水準 の和 = 固有 値の和 = 0.2577]	

回答者のクラスター構成 (= 記述した回答)

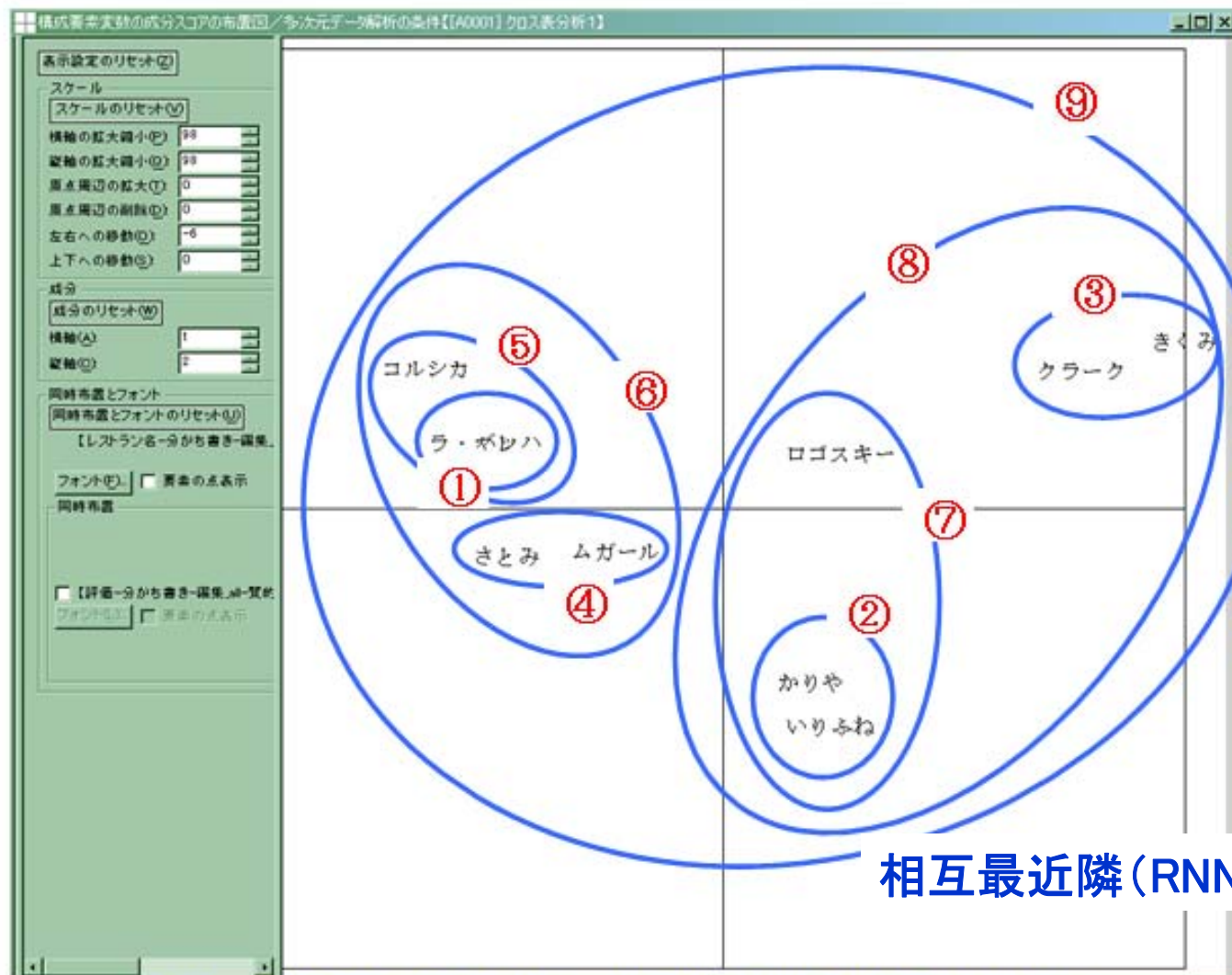
レストランのクラスター構成

この階層の結合水準値の総和 =
総変動 (固有値総和, 慣性)

デンドログラム, 結合情報を確認



クラスター化過程の確認



①
↓
②
↓
③
↓
⋮
↓
⑨と進む
ボトムアップに併合する

相互最近隣(RNN)の規則の確認

図 8 階層的分類のクラスター化過程, 入れ子構造のイメージ

クラスター化過程の要約を再確認

結合の ステップ (r)	クラスター数 (g)	階層水準に 含まれる レストラン数 (t)	階層水準に 含まれる 回答者数 (n)	併合先の群 (C_p)	併合された群 (C_q)	階層の 結合水準値 $h(r)$	階層の 結合水準値の累積和 $\sum_r h(r)$
①	9	2	288	バッハ	ラ・マレ	0.0003	0.0003
②	8	2	333	いりふね	かりや	0.0005	0.0008
③	7	2	212	きくみ	クラーケ	0.0016	0.0024
④	6	2	204	さどみ	ムガール	0.0017	0.0041
⑤	5	3	410	ゴルシカ	バッハ (バッハUラ・マ レ)	0.0022	0.0063
⑥	4	5	614	さどみ	ゴルシカ (ゴルシカUバッハ Uラ・マレ)	0.0097	0.0160
⑦	3	3	458	いりふね (いりふねUかりや)	ロゴスキー	0.0154	0.0314
⑧	2	5	670	いりふね (いりふねUかりやU ロゴスキー)	きくみ (きくみUクラー ケ)	0.0679	0.0993
⑨	1	10	1,284	いりふね (いりふねUかりやU ロゴスキーUきくみU クラーケ)	さどみ (さどみUゴルシカ UバッハUラ・マ レ)	0.1584	0.2577
						0.2577 [結合水準 の和 = 固有 値の和 = 0.2577]	

たとえば, 1つの併合をみると, ...

結合の ステップ (r)	クラスター数 (g)	階層水準に 含まれる レストラン数 ($+$)	階層水準に 含まれる 回答者数 ($+$)	併合先の群 (C_p)	併合された群 (C_q)	階層の 結合水準値 $h(r)$	階層の 結合水準値の累積和 $\sum_r h(r)$
①	9	2	288	バッハ	ラ・マレ	0.0003	0.0003
②	8	2	333	いりふね	かりや	0.0005	0.0008
③	7	2	212	きくみ	クラーク	0.0016	0.0024
④	6	2	204	さどみ	ムガール	0.0017	0.0041
⑤	5	3	410	ゴルシカ	バッハ (バッハUラ・マ レ)	0.0022	0.0063
C_p と C_q が併合を $C_p \triangleq C_p \cup C_q$ と書く (p と q とが併合し, あらたにクラスターとする) $C_p = \{\text{バッハ}\}, C_q = \{\text{ラ・マレ}\}$ \Downarrow 併合して $C_p \cup C_q = \{\text{バッハ}, \text{ラ・マレ}\} \triangleq \{\text{バッハ}\}$					ゴルシカ (ゴルシカUバッハ Uラ・マレ)	0.0097	0.0160
					ロゴスキー	0.0154	0.0314
					きくみ (きくみUクラー ク)	0.0679	0.0993
					さどみ (さどみUGオルシ カUバッハUラ・マ レ)	0.1584	0.2577
						0.2577 [結合水準 の和=固有 値の和= 0.2577]	

ここで何を行うのか？

- ここでこの“階層的分類は何を行った”のだろうか．あるいはこれで“なにを調べよう”としたのか．
- 2元クロス表の行側(レストラン)の分類を例とするが，これは同時に，それを選んだ回答者(行和)の分類でもあることに注意する．
- ここで“成分スコア”はどう使われたのか．
- つまりカイ二乗距離とユークリッド距離はどう機能し，用いられたのか．
- カイ二乗統計量あるいは総変動(全慣性)はどう用いるのか．

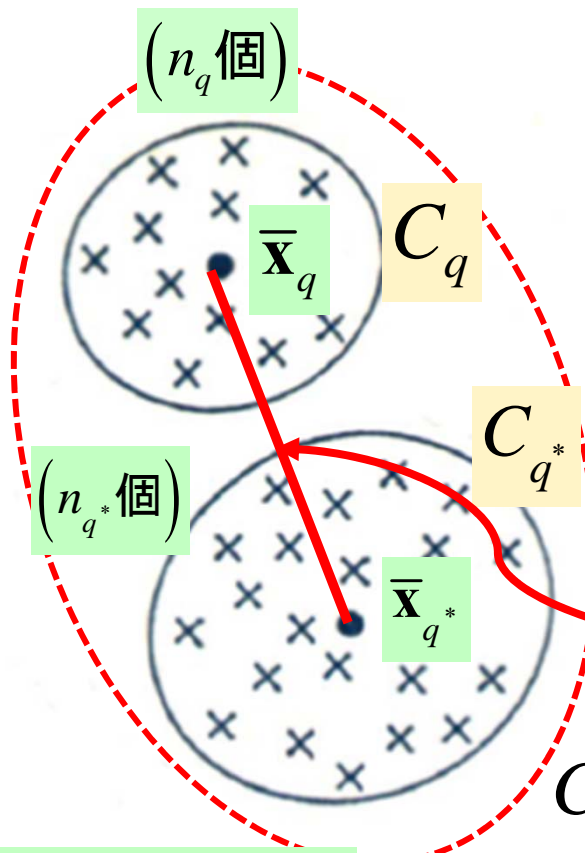
(つづき)

- まず, 成分スコアに“ユークリッド距離”を適用し“ワード基準”による階層分類を行う仕組みを述べる.
- この分類操作をレスランデータで調べる.
- 次に, “カイ二乗統計量の分解・併合”による階層分類を調べる. 総変動(全慣性)の分解でもある.
- 実は, これら両者は同じ操作を行っていること.
- 操作手順を比較し, 何を行っているかを確認する.

ワード基準とその考え方

- 一般に、階層的分類法では、2つのクラスターの併合による平方和の変化(増分)を距離と考える.
- この小さいクラスターから併合を順に進める. つまり階層構造を作る.
- ウォード法は、ワード基準を用いる(J. Wardの提案した方法).
- 平方和の分解, つまりデータの変動を平方和で測り, クラスターの併合時の距離とする. これを簡単に述べる.
- “平方和”ということは“平方ユークリッド距離”を用いることに同じことに注意.

2つのクラスターとその併合(図解)



2つのクラスター C_q と C_{q^*} の併合: $C_q \cup C_{q^*}$

あらたにできるクラスター: $C_t = C_q \cup C_{q^*}$

このときの変動の変化を平方和で測る.

実際は, $C_q \cup C_{q^*}$ の併合は, q^* は q に吸収(併合)されたと考える(ラベリング処理). $C_q \triangleq C_q \cup C_{q^*}$

重心間の重み付き平方距離

$$d_w^2(q, q^*) \triangleq \frac{n_q n_{q^*}}{n_q + n_{q^*}} \left\| \bar{\mathbf{x}}_q - \bar{\mathbf{x}}_{q^*} \right\|^2$$

$$C_t = C_q \cup C_{q^*} \Rightarrow C_q$$

$\bar{\mathbf{x}}_q, \bar{\mathbf{x}}_{q^*}$ は, 重心ベクトル

- このとき, ウォード基準とは次の関係を用いること



2つのクラスター併合時の変動の関係

$$J_t = \underbrace{J_q + J_{q^*}}_{\text{(2つのクラスターの平方和の和)}} + \underbrace{\frac{n_q n_{q^*}}{n_q + n_{q^*}} \left\| \bar{\mathbf{x}}_q - \bar{\mathbf{x}}_{q^*} \right\|^2}_{\text{(併合によって増えた変動の増分)}} \quad (\star)$$

$$\left\{ \begin{array}{l} \text{クラスター } C_q \text{ の平方和: } J_q \\ \text{クラスター } C_{q^*} \text{ の平方和: } J_{q^*} \\ \text{級内変動: } J_q + J_{q^*} \text{ (2つのクラスター } C_q, C_{q^*} \text{ の平方和の和)} \\ \text{2つのクラスター間の級間変動: } \frac{n_q n_{q^*}}{n_q + n_{q^*}} \left\| \bar{\mathbf{x}}_q - \bar{\mathbf{x}}_{q^*} \right\|^2 \quad (\star\star) \end{array} \right.$$

- ここで「級内変動」＝「クラスター内変動」のこと.
- 「級間変動」＝「クラスター間変動」のこと.

さらに, ...

2つのクラスター C_q, C_{q^*} 間の重心間距離: $\|\bar{\mathbf{x}}_q - \bar{\mathbf{x}}_{q^*}\|^2$

ここで, $\bar{\mathbf{x}}_q$ はクラスター C_q の重心ベクトル,

$\bar{\mathbf{x}}_{q^*}$ はクラスター C_{q^*} の重心ベクトル

n_q : クラスター C_q のクラスター・サイズ

n_{q^*} : クラスター C_{q^*} のクラスター・サイズ

$n_t = n_q + n_{q^*}$: 併合クラスター $C_t = C_q \cup C_{q^*}$ のクラスター・サイズ

- ここで, 式(★★)の加重は, 以下のように, 2つのクラスター・サイズの“調和平均”になっている.

$$\frac{1}{\frac{1}{n_q} + \frac{1}{n_{q^*}}} = \frac{n_q n_{q^*}}{n_q + n_{q^*}} \quad (★★★)$$

クラスター化(併合)が進むと, ...

- (2つの)クラスターの併合を考えたとき, “距離が近い個体あるいはクラスター”から, 併合する, とする.
- 平方和を使うことは, 平方ユークリッド距離を用いることに同じであったことを想起.
- “級間変動”(クラスター間変動)が小さい個体またはクラスターを併合するというルールに従うこと.
- これは“(2つの)クラスター間距離”として, 以下を考えることに同じである(重み付きの平方ユークリッド距離).

$$\text{2つのクラスター } C_q \text{ と } C_{q^*} \text{ の距離: } d_w^2(q, q^*) \triangleq \frac{n_q n_{q^*}}{n_q + n_{q^*}} \left\| \bar{\mathbf{x}}_q - \bar{\mathbf{x}}_{q^*} \right\|^2$$

(クラスターの重心間平方距離に加重を付けたもの)

これを「成分スコア」に適用する

- このワード基準を用いる併合を，対応分析で得た成分スコアに適用する.
- (元のクロス表のストレッチ・プロファイル間の)カイ二乗距離は，成分スコアのユークリッド距離であった(何度も指摘した).
- よって，成分スコアにワード基準を用いることは妥当かつ自然な発想である.

(つづき)

- 注意することは、式(★★★)の“加重”をどう考えるかである.
- クロス表とプロフィールを考えると、“質量”を用いて加重化することが自然である.
- たとえば、レストラン分類は、あるレストランを選んだ回答者(数)の重みが付いている、その回答者の分類でもある.
- これを考慮して、次ページのクラスター間距離を考える.

対応分析におけるワード基準

- ここまでの説明から, 以下のように考える([スライド83](#)参照).

$$d_w^2(q, q^*) = \frac{m_q m_{q^*}}{m_q + m_{q^*}} \left\| \bar{\mathbf{z}}_q - \bar{\mathbf{z}}_{q^*} \right\|^2$$

ここで, m_q, m_{q^*} は, 2つのクラスター C_q, C_{q^*} それぞれの質量

$$m_q = \frac{f_{q+}}{N} (\text{クラスター } C_q \text{ の質量}), \quad m_{q^*} = \frac{f_{q^*+}}{N} (\text{クラスター } C_{q^*} \text{ の質量})$$

また, $\bar{\mathbf{z}}_q, \bar{\mathbf{z}}_{q^*}$ は以下

$$\bar{\mathbf{z}}_q, \bar{\mathbf{z}}_{q^*} \left(\begin{array}{l} \text{2つのクラスター } C_q, C_{q^*} \text{ の重心座標ベクトル} \\ \text{つまり, クラスター } C_q, C_{q^*} \text{ の“重心の成分スコア”} \end{array} \right)$$

さらに、書き替えて確認

- 成分スコア・ベクトルの要素単位で書き替えてみる.
- ここで、全成分数(K)としたが、レストラン・データの場合は、 $K=2$ と(2成分)すればよい.
- つまり、2次元の布置図の中での、個体間あるいはクラスター間の重み付き(平方)ユークリッド距離となる.

$$d_W^2(q, q^*) = \frac{m_q m_{q^*}}{m_q + m_{q^*}} \left\| \bar{\mathbf{z}}_q - \bar{\mathbf{z}}_{q^*} \right\|^2 \Rightarrow d_W^2(q, q^*) = \frac{m_q m_{q^*}}{m_q + m_{q^*}} \sum_{k=1}^K \left(\bar{z}_{qk} - \bar{z}_{q^*k} \right)^2$$

レストラン・データの場合(行側, レストランの分類), $K=2$ として以下

$$d_W^2(q, q^*) = \frac{m_q m_{q^*}}{m_q + m_{q^*}} \sum_{k=1}^2 \left(\bar{z}_{qk} - \bar{z}_{q^*k} \right)^2 \quad (★★★★)$$

ここで注意すること

- 階層的分類は，初期設定は，個々の分類対象を“個々の単一のクラスター”とみなす.
- レストランの分類であれば，10のレストランが10個のクラスターと考える．（初期化）
- こうしたクラスターサイズが「1」のときを“シングルトン”という.
- 出発時は，クラスターサイズはすべて1であるから，級内変動（平方和の和）は「0」である（変動，バラツキはない）.
- 重心の成分スコアは，そのレストランに対する成分スコアそのものである（初期化）.

(つづき)

- 前に確認したデンドログラムと結合順序(階層化)情報から, 段階を追って, レストランの併合, 分類を行ってみる.
- 第1ステップ(はじめの併合)は, 「バッハ」と「ラ・マレ」であった.
- 第2ステップは, 「いりふね」と「かりや」の併合となる.
- 以下, 順に, 9回の併合を繰り返すと, 最後は全レストラン(と全回答者)が「1つのクラスター」にグループ化され, これで終了.

(つづき)

- ここまでにいくつかのスライドで示した“**クラスター化過程**”の一部を数値例として確認する.
- 単なる数値確認であるから, ざっとみる. 要点は以下.
- レストランあるいはクラスターの併合で, クラスターの重心が変わる(重心の成分スコアの“更新”).
- クラスター併合時の距離(ワード距離)を求める.
- 第1ステップと第5ステップを調べる.

第1ステップ(①)

$$C_q = \{\text{バッハ}\}, C_{q^*} = \{\text{ラ・マレ}\}$$

↓ この2つが併合して次のクラスターが生成される



$$C_q \cup C_{q^*} = \{\text{バッハ}, \text{ラ・マレ}\} \Leftrightarrow \text{これがあらたな}\{\text{バッハ}\}\text{というクラスター}$$

クラスター [↗]	クラスター内変動 [↗]	レストラン数 [↗]	回答者数 [↗]	周辺確率 [↗] (質量) [↗]	成分スコア 1 [↗] (重心スコア) [↗]	成分スコア 2 [↗] (重心スコア) [↗]
バッハ: C_q (シングルトン) [↗]	0 [↗]	1 [↗]	142 [↗]	0.1106 [↗] (m_q) [↗]	-0.3966 [↗]	0.1220 [↗]
ラ・マレ: C_{q^*} (シングルトン) [↗]	0 [↗]	1 [↗]	146 [↗]	0.1137 [↗] (m_{q^*}) [↗]	-0.4636 [↗]	0.1191 [↗]
↓ [↗]						
バッハ ∪ ラ・マレ [↗] $C_q \cup C_{q^*}$ [↗]	0.0003 [↗]	2 [↗]	288 [↗]	0.2243 [↗] ($m_q + m_{q^*}$) [↗]	-0.4305 [↗]	0.1205 [↗]

- 併合が起きた箇所の情報を切り出して要約した.
- この表から, 前述の式(★★★★)から距離を算出.

併合距離を算出

$$d_w^2(q, q^*) = \frac{m_q m_{q^*}}{m_q + m_{q^*}} \sum_{k=1}^2 (z_{qk} - z_{q^*k})^2$$

↓

$$\begin{aligned} & \frac{0.1106 \times 0.1137}{0.2243} \times \left\{ [-0.3966 - (-0.4636)]^2 + [0.1220 - 0.1191]^2 \right\} \\ &= 0.056064288 \times (0.004489 + 0.00000841) = 0.000252144 \doteq \underline{0.000252} \quad (\blacktriangledown) \end{aligned}$$

併合後のクラスターの重心成分スコアの更新

$$\text{成分スコア1の重心スコア: } \frac{m_q z_{q1} + m_{q^*} z_{q^*1}}{m_q + m_{q^*}} = \frac{\{0.1106 \times (-0.3966) + 0.1137 \times (-0.4636)\}}{0.2243} = \underline{\underline{-0.4305}}$$

$$\text{成分スコア2の重心スコア: } \frac{m_q z_{q2} + m_{q^*} z_{q^*2}}{m_q + m_{q^*}} = \frac{\{0.1106 \times 0.1220 + 0.1137 \times 0.1191\}}{0.2243} = \underline{\underline{0.1205}}$$

- 以上が前ページの表に要約されている。

第5ステップ(⑤)

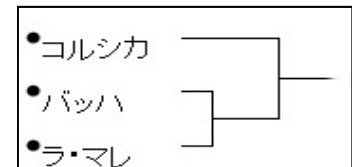
$$C_q = \{\text{バッハ, ラ・マレ}\}, C_{q^*} = \{\text{コルシカ}\} \Rightarrow C_q \cup C_{q^*} = \{\text{バッハ, ラ・マレ, コルシカ}\}$$

クラスター [↙]	クラスター内変動 [↙]	レストラン数 [↙]	回答者数 [↙]	周辺確率 [↙] (質量) [↙] (m_q) [↙]	成分スコア 1 [↙] (重心スコア) [↙]	成分スコア 2 [↙] (重心スコア) [↙]
バッハ: C_q [↙] (バッハ ∪ ラ・マレ) [↙]	0.0003 [↙]	2 [↙]	288 [↙]	0.2243 [↙] (m_q) [↙]	-0.4305 [↙]	0.1205 [↙]
コルシカ: C_{q^*} [↙]	0 [↙]	1 [↙]	122 [↙]	0.0950 [↙] (m_{q^*}) [↙]	-0.5497 [↙]	0.2586 [↙]
⇓ [↙]						
バッハ ∪ コルシカ [↙] $C_q \cup C_{q^*}$ [↙]	0.0025 [↙]	3 [↙]	410 [↙]	0.3193 [↙] (m_{q^*}) [↙]	-0.4660 [↙]	0.1616 [↙]

$$d_w^2(q, q^*) = \frac{0.2243 \times 0.0950}{0.3193} \times \left\{ [-0.4305 - (-0.5497)]^2 + [0.1205 - 0.2586]^2 \right\}$$

$$= 0.066735045 \times (0.01420864 + 0.01907161) = 0.002220958 \doteq 0.00222$$

併合後のクラスターの重心成分スコアの更新



成分スコア1の重心スコア: $\frac{m_q z_{q1} + m_{q^*} z_{q^*1}}{m_q + m_{q^*}} = \frac{\{0.2243 \times (-0.4305) + 0.0950 \times (-0.5497)\}}{0.3193} = -0.466596 \dots \doteq -0.4660$
.....

再度，クラスター化過程を確認

ステップ (r)	クラスター数 (g)	階層水準に 含まれる 回答者数	併合先の群 (C_q)	併合された群 (C_{q^*})	階層の 結合水準値 $h(r)$
①	9	288	バッハ	ラ・マレ	(▼) 0.00025
②	8	333	いりふね	かりや	0.00052
③	7	212	きくみ	クラーク	0.00163
④	6	204	さとみ	ムガール	0.00165
⑤	5	410	コルシカ	バッハ (バッハ ∪ ラ・マレ)	(▼) 0.00222
⑥	4	614	さとみ	コルシカ (コルシカ ∪ バッハ ∪ ラ・マレ)	0.00973
⑦	3	458	いりふね (いりふね ∪ かりや)	ロゴスキー	0.01538
⑧	2	670	いりふね (いりふね ∪ かりや ∪ ロゴスキー)	きくみ (きくみ ∪ クラーク)	0.06788
⑨	1	1,284	いりふね (いりふね ∪ かりや ∪ ロゴスキー ∪ きく み ∪ クラーク)	さとみ (さとみ ∪ コルシカ ∪ バッハ ∪ ラ・マレ)	0.15842
(※) もとのデータ表からみた，つまり回答者が選んだレストラン数による計数．最後のセルが 1,284 名の回答者に相当，表 3 を参照．					0.25768 [結合水準の和 = 固有値の和]

(赤文字の箇所を数値チェックで確かめた)

階層の結合水準値とクラスター間距離

- 「階層の結合水準値」とワード基準の(2つの)「クラスター間距離」との関係は重要である.

ステップ「 r 」における結合水準: $h(r)$

$$h(r) \equiv d_E^2(q, q^*) = \frac{m_q m_{q^*}}{m_q + m_{q^*}} \sum_{k=1}^2 (z_{qk} - z_{q^*k})^2$$

- [階層の結合水準値の和] = [固有値の和・総変動(全慣性)]

$$\sum_r h(r) = \sum_k \lambda_k$$

もう一度確認すると, ...

- 確かに, 「階層の結合水準値」と「クラスター間距離」, そして「固有値の和・総変動(全慣性)」の関係が確認できる.

結合の ステップ (r)	クラスター数 (g)	階層水準に 含まれる レストラン数 (t)	階層水準に含まれる 回答者数 (⊕)	併合先の群 (C_p)	併合された群 (C_q)	階層の 結合水準値 $h(r)$	階層の結合水準値 の累積和 $\sum_r h(r)$
①	9	2	288	バッハ	ラ・マレ	0.0003	0.0003
②	8	2	333	いりふね	かりや	0.0005	0.0008
③	7	2	212	きくみ	クラーケ	0.0016	0.0024
④	6	2	204	さどみ	ムガール	0.0017	0.0041
⑤	5	3	410	ゴルシカ	バッハ	0.0022	0.0063
⑥	4	5	614	さどみ	ゴルシカ	0.0097	0.0160
⑦	3	3	458	いりふね	ロゴスキー	0.0154	0.0314
⑧	2	5	670	いりふね	きくみ	0.0679	0.0993
⑨	1	10	1,284	いりふね	さどみ	0.1584	0.2577
(⊕) クラスター内に入ったレストラン数 (最後のセルが 10 クラスター) (⊕) もとのデータ表からみた, つまり回答者が選んだレストラン数による計数 (最後のセルが 1,284 名の回答者に相当), 表 3 を参照.						0.2577 [結合水準 の和=固有 値の和= 0.2577]	

「カイ二乗統計量の分解」からみた分類

- 成分スコアを用いた(平方)ユークリッド距離によるワード基準のクラスター化を調べた.
- これを, 元の2元クロス表の総変動(全慣性)の分解による階層分類に読み替えてみる.
- データが与えられると“ある一定の値”が“総変動”として得られること(データ表の全情報).
- これを“クラスター間変動(級内変動)”と“クラスター内変動(級内変動)の和”に分解することを考える.

平方和分解・二乗和分解については統計学の関連書を参照. 基本的かつ重要な操作. とくに「分散分析」で用いる.

(つづき)

- これは. いわゆる“平方和の分解”を用いること.
- 対応分析では“ホイヘンスの分解公式”(Huyghens decomposition)という.
- [総変動(全慣性)] = [クラスター間変動] + [クラスター内変動の和]
- これと[総変動(全慣性)] = [固有値の総和] = [ピアソンのカイ二乗統計量 / 総度数] の関係を見る.
- これらの関係を確認する. 記法の説明は順に.

$$S_T = \underbrace{S_B(g)}_{\text{クラスター間変動}} + \underbrace{\sum_{l=1}^g S_W(g, l)}_{\text{クラスター内変動の和}} \Leftrightarrow S_T = \frac{\chi_p^2}{N} (= \phi^2)$$

成分スコア空間(合成指標)

\Leftrightarrow プロファイル空間

ホイヘンス(Huyghens)はホイゲンスということもある.

クラスターの特徴を表す統計量

- この仕組み(統計量の相互の関係)を用いて, どのようなクラスターが作られたかを統計量で評価する.
- 基本は, 全情報である“総変動”を, 分割でえたクラスターの“クラスター内変動(の和)”と“クラスター間変動”に分けて評価すること(分解すること).
- この考え方は, 統計のもっとも基本的な操作の1つである, 全変動(全平方和や全分散)を複数の変動の要素に分ける“平方和分解”(二乗和分解)の仕組みに同じこと.

(つづき)

- データ(点, 成分スコア)の示す変動という情報に注目することで, データの特徴抽出を行えるだろうという考え方.
- 例でみたように, データがいったん与えられると, “全変動”は一定となり(確定), これを“クラスター間変動”と“クラスター内変動”が分け合うことになる.
- なにごとも万能ではない, この統計量・指標を用いることは, これに特有の性質の影響がある. 知って用いる.
- たとえば, はずれ値の影響を受けやすい. ワイルドショット(的外れ=重心がずれたクラスター)を作る傾向など.

特徴, 注意すること, 対応策(考え方)

- 基本は“クラスターの等質性”あるいは“まとまり具合”を平方和のような統計量で測る, ということ. (†)
- ウォード法, k -平均法の説明でみたように, “平方和や分散”つまり“平方ユークリッド距離”で, クラスター化の程度, まとまりの良さを測る.
- よって, はずれ値の影響を受けやすい.
- これに対する対策として, “ハイブリッド法”を使う.
- WordMinerでは, ウォード法と k -平均法との併用でクラスター化をチューニングする.

(†) 平方和(あるいは分散)を使うから, 見た目が“まとまりのよい”クラスターが作られる, というのが正しい.
喩え: フライパンに油の球.

(つづき)

- (ワード法による)初期分類を行う.
- この不具合を(例:はずれ値の影響などでそれが生じたなら)再配置と再分類(細分類)でクラスターのまとまり具合を調整する.
- 対応分析ははずれ値が生じやすいから,この手当は効果的に機能する.
- 平方和・分散型で測れるようなクラスターの“存在”は稀であるから(≡目立ったクラスターが見えない),この対策は,比較的うまく機能する可能性.

各統計量は何を測るのか(要点のみ)

- 登場する統計量は, それぞれデータ(成分スコア)のどのような特徴を測っているのか.
- 式の具体的な記述・説明はせずに, それぞれが何を意味するか, 要点を述べる.
- とくに, 例としてクラスター数(g)を“5群”と指定したときに, 各統計量をどう読むか, をざっと調べる.
- おもに, 以下についてふれる
 - “クラスター数”の考え方(実は, 最適解がないこと)
 - 得られた“クラスターの解釈”(ある種の検定操作)
 - これは, 重心の成分スコアを評価すること
- その他, さまざまな課題があるが, ここでは取り上げない.

クラスターの特徴を測る(記法の準備)

- ここで, 以下のような記号, 記法を用意する.
- 個々の式の細かい誘導は行わない. 意味を知ること.
- ここで次の[性質5]を確認する.

クラスター数: g (これは階層の水準に対応することに注意)

総変動: S_T (あらためてこの記号で表す)

クラスター内変動: $S_W(g, l)$ (クラスターを g 群としたときの, そこに含まれるあるクラスター l の郡内変動, よって $l = 1, 2, \dots, g$)

クラスター内変動の和: $\sum_{l=1}^g S_W(g, l)$

クラスター間変動: $S_B(g)$ (クラスターを g 群としたときの群間変動)

クラスター間変動比: $\eta_g = \frac{S_B(g)}{S_T} \times 100$ (%) (g 群のとき)

※これらの記号の要約は, テキスト, 第Ⅲ部, 9ページあたり

クラスターの特性を表す統計量

[性質5] 文字と式で書けば以下の関係にある.

$$\begin{aligned} \text{[総変動]} &= \text{[固有値の和つまり全慣性]} \\ &= \text{[クラスター間変動]} + \text{[クラスター内変動の和]} \end{aligned}$$

$$\begin{aligned} \text{[全慣性: Total inertia]} \\ &= \text{[群間慣性: between-clusters inertia]} \\ &\quad + \text{[群内慣性: within-clusters inertia]} \end{aligned}$$

$$S_T = \frac{\chi_p^2}{N} \left(= \sum_{k=1}^K \lambda_k \right) \quad (\text{何度も確認した関係})$$

⇕

$$S_T = S_B(g) + \sum_{l=1}^g S_W(g, l) \quad (\text{クラスターの変動の関係式})$$

目標は, ...

- ① [総変動]: データが与えられると, ある“一定の値”として決まる(情報の全体).
- ② [クラスター間変動]: クラスター間はなるべく離れるように, つまり大きくしたい[分離度, 乖離度の改善]
- ③ [クラスター内変動]: 個々のクラスター内変動はなるべく小さくしたい(等質でそろっているように).
- ④ [クラスター内変動の“和”]: ③が(ある程度)満たされればこれも小さくなるだろう.
- ⑤ ①を満たしながら, ②と③を“最適化”する分割とすること.

つまり, ...

- 右辺の2つの項はトレードオフの関係にある.
- 対応分析の世界だけでなく“統計学一般の基本的な考え方”である.
- とくに, クラスター化を考えると時の基本原理となっている.

$$\begin{array}{ccc} S_T & = & S_B(g) \quad + \quad \sum_{l=1}^g S_W(g, l) \\ \Downarrow & & \Downarrow \qquad \qquad \qquad \Downarrow \\ \left[\text{一定} \right] & \left[\begin{array}{c} \text{なるべく大きく} \\ \text{離したい} \end{array} \right] & \left[\begin{array}{c} \text{なるべく小さく} \\ \text{まとめたい} \end{array} \right] \\ & \text{クラスター内変動} & \text{クラスター間変動} \end{array}$$

もう1つ統計量を用意

- ある階層水準における, つまりあるクラスター数を指定したときのカイ二乗統計量を考える.
- これは, クロス表の圧縮化の反復, つまりカイ二乗統計量 (とカイ二乗距離) を用いたクラスター化過程で得られる “カイ二乗統計量の分解” に相当する.
- クラスター数「 g 」のときの, カイ二乗統計量を以下で表す.
- レストランを, $g=10$ (群) つまり未分類のときをクラスター・サイズ1の10個のクラスターと考えること.

$$\chi_p^2(g) \left(\begin{array}{l} \text{ある階層水準, クラスター数}(g) \\ \text{におけるカイ二乗統計量} \end{array} \right)$$
$$g=10 \text{ のとき (元のクロス表), } \chi_p^2(10) \equiv \chi_p^2$$

カイ二乗統計量は分解できるという性質がある, これを言い換えると“加法性”があるという.

もとのクロス表と総変動ほかの関係(重要)

	<はじめのクロス表>	<統計量と生成される圧縮化クロス表の履歴>		
	(*) 10 群としたことに相当	工夫・サービス	味	量
10 群	いりふね	98	25	32
	かりや	105	35	38
	きくみ	35	8	67
	さとみ	42	46	7
	クラーク	34	14	54
	コルシカ	32	77	13
	パツハ	48	76	18
	ムガール	49	44	16
	ラ・マレ	49	82	15
	ロゴスキー	48	35	42
	↓	↓		
	$\chi_p^2 = \chi_p^2(10) = 330.860$ (クロス表から得たカイ二乗統計量)	総変動 (分類前) 0.2577	$0.257679 \times 1284 =$ 330.860	クラス-内変動の 和 = 0

$$\chi_p^2(g)$$

(ある階層水準, クラスタ数(g))
 におけるカイ二乗統計量

テキスト, 表20(34ページあたり)

考え方

- クラスター化前のクロス表, これを“個々のレストランがサイズ1の10個のクラスター”と考える(各クラスター内変動=0).
- ここで, 「総変動(全慣性) = 固有値の総和」(= $0.2577 \div 0.2576$) に注意する. この値は一定値のまま.
- ここから, レストラン(かつ回答者)のクラスター化を開始する.
- 9群から6群までは省略し, 続く5群以下の併合を確認する.
- 5群. 4群は, レストランの分類によって“クロス表が圧縮された”(クラスター化でレストランが併合)と考える.
- カイ二乗統計量が変化する(情報が次第に減る).

5群 ($g=5$) から4群 ($g=4$) へ

<5 群に分類後の 併合を以下で追跡>

クラスター化履歴		工夫・サービス		味	量
5 群	{さとみ, ムガール}	91	90		23
	{パッパ, コルシカ, ラ・マレ}	129	235		46
	{ロゴスキー}	48	35		42
	{いりふね, かりや}	203	60		70
	{クラーク, きくみ}	69	22		121

⇓

$$\chi_p^2(5) = 322.799$$

クラスター間変動
 $S_B(5) = 0.2514$

$$0.251401 \times 1284 = 322.799$$

クラスター内変動の
和 = 0.0062

4 群	{さとみ, パッパ, ムガール, コルシカ, ラ・マレ}	220	325	69
	{ロゴスキー}	48	35	42
	{いりふね, かりや}	203	60	70
	{クラーク, きくみ}	69	22	121

⇓

$$\chi_p^2(4) = 310.308$$

クラスター間変動
 $S_B(4) = 0.2417$

$$0.241673 \times 1284 = 310.308$$

クラスター内変動の
和 = 0.0159

0.2514 + 0.0062 = 0.2576

ここで、クラスター間変動は、その群におけるクロス表に対応分析を適用したときの総変動(固有値の総和)となっている。

3群 ($g=3$) から1群 ($g=1$) へ

3 群	{さとみ, バッハ, ムガール, コルシカ, ラ・マレ}	220	325	69
	{いりふね, ロゴスキー, かりや}	251	95	112
	{クラーク, きくみ}	69	22	121
⇓		⇓		
$\chi^2_p(3) = 290.564$		クラスター間変動 $S_B(3) = 0.2263$	0.226296×1284 $= 290.564$	クラスター内変動の 和 $= 0.0313$
2 群	{さとみ, バッハ, ムガール, コルシカ, ラ・マレ}	220	325	69
	{いりふね, クラーク, ロゴスキー, きくみ, かりや}	320	117	233
⇓		⇓		
$\chi^2_p(2) = 203.405$		クラスター間変動 $S_B(2) = 0.1584$	0.158415×1284 $= 203.405$	クラスター内変動の 和 $= 0.0992$
1 群	{さとみ, バッハ, ムガール, コルシカ, ラ・マレ}	540	442	302
	{いりふね, クラーク, ロゴスキー, きくみ, かりや}			
$\chi^2_p(1) = 0.0$		□	+	□
		$= 0.2576$		

ここで“クラスター間変動”と“クラスター内変動の和”とある統計量はなにか？

一覧に要約(テキスト, 33ページ, 表19)

- クラスター数が増えると, “クラスター間変動”(①欄)は単調に増え, “クラスター内変動の和”(④)は単調に減る.
- 両者を加えると(①+④)は“総変動・全慣性”(固有値の総和), これは「一定の値」である.

ステップ (r)	クラスター数 (g)	① クラスター間変動 $S_B(g)$	② = ① × 1,284 (s) カイニ乗統計量 に相当 $\chi_p^2(g)$	③クロス表から 算出のとき カイニ乗統計量 $\chi_p^2(g)$	④ クラスター内 変動の和 $\sum_{l=1}^g S_w(g, l)$	⑤チェック ①+④ S_T [総変動 = 固有値の和]
⑧	2	0.1584	203.4049	203.405	0.0992	0.2576
⑦	3	0.2263	290.5692	290.564	0.0313	0.2576
⑥	4	0.2417	310.3043	310.308	0.0159	0.2576
⑤	5	0.2514	322.7976	322.799	0.0062	0.2576
④	6	0.2536	325.6481	325.648	0.0040	0.2576
③	7	0.2553	327.7667	327.767	0.0024	0.2577
②	8	0.2569	329.8596	329.860	0.0008	0.2577
①	9	0.2574	330.5401	330.540	0.0003	0.2577
初期	10	0.2577	330.8598	330.860	0.0000	0.2577

① 単調に増える

④ 単調に減る

④ 単調に増える

一定の値(総変動)

$$\sum_{k=1} \lambda_k = \frac{\chi_p^2(g)}{N}$$

別の要約表を観察(テキスト30ページ, 表17相当)

ステップ (r)	クラスター 数 (g)	①階層水準に 含まれる 異なり構成要素 数	②階層水準に 含まれる 構成要素数	③階層の 結合水準 $h(r)$	④結合水準の 累積和 $\sum_r h(r)$	⑤総変動に占め る割合(%)	⑤デンドログラムの高のさの比
	クラスター の遷移	(a) クラスター に含まれるレス トランの数	(b) クラスター 内のサンプル 数	(c) デンドロ グラムで確認	(d) 各水準のクラ スター内変動の和	③÷総変動(固 有値総和)×100 (%)	
①	9	2	288	0.00025	0.00025	0.10	
②	8	2	333	0.00052	0.00077	0.20	
③	7	2	212	0.00163	0.00240	0.63	
④	6	2	204	0.00165	0.00405	0.64	
⑤	5	3	410	0.00222	0.00627	0.86	
⑥	4	5	614	0.00973	0.01600	3.78	
⑦	3	3	458	0.01538	0.03138	5.97	
⑧	2	5	670	0.06788	0.09926	26.34	
⑨	1	10	1284	0.15842	0.25768 [固有値の和]	61.48	
—	—	—	—	0.25768 [結合水準の 和]	← ($\lambda_1 + \lambda_2$)	100.00	

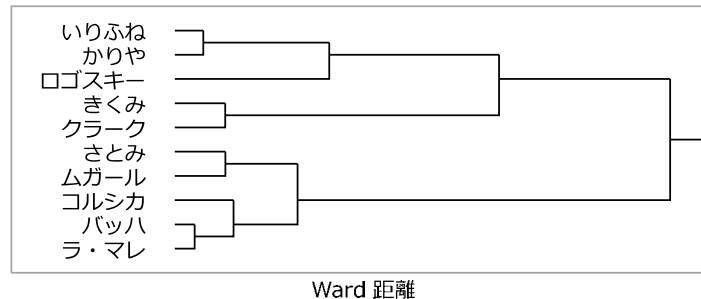
総変動(全慣性) = 固有値の和

(前ページの点線枠に同じ値) 7

結果をJMPスクリプトで確認

- JMPスクリプトの対応分析で, デンドログラムと結合距離
他の統計量を確認する.
- 「ワード距離の2乗」=階層の結合水準 $h(r)$ (表の③欄)
に相当.

樹形図



クラスター分析履歴

クラスター数	Ward 距離	Ward 距離の2乗	併合先の群	併合された群	併合先の周辺割合(%)	併合された周辺割合(%)
9	0.01588	0.00025	バッハ	ラ・マレ	11.1	11.4
8	0.02287	0.00052	いりふね	かりや	12.1	13.9
7	0.04042	0.00163	きくみ	クラーク	8.6	7.9
6	0.0406	0.00165	さとみ	ムガール	7.4	8.5
5	0.04712	0.00222	コルシカ	バッハ	9.5	22.4
4	0.09863	0.00973	さとみ	コルシカ	15.9	31.9
3	0.12401	0.01538	いりふね	ロゴスキー	25.9	9.7
2	0.26054	0.06788	いりふね	きくみ	35.7	16.5
1	0.39801	0.15842	いりふね	さとみ	52.2	47.8

④単調に増える

Ward 距離2乗の合計 0.257678804157674

最終確認: 数値の見方, 関係

- ここで, “階層の結合水準”(③)は(2つの)クラスター併合時の高さ(併合時の距離)に相当する.
- 前にみたデンドログラムと表で確認しよう.
- “結合水準の累積和”(④)は, “クラスター内変動の和”の変化(桁数が違うが同じ情報). 各表で比べてみよう.

(重要な関係)

$$S_T = S_B(g) + \sum_{l=1}^g S_W(g, l)$$

⇕

(総変動) = (クラスター間変動) + (クラスター内変動の和)

$$S_T = \sum_{k=1}^K \lambda_k = \frac{\chi_p^2}{N}$$

⇕

(総変動) = (固有値の和)
= (ピアソンのカイ二乗統計量 / 総度数)

$$\sum_r h(r) = \sum_k \lambda_k = \frac{\chi_p^2}{N}$$

(結合水準値の和) = (固有値の和)

階層結合水準, 成分数, クラスター数の目安

[性質4]

- クラスター化で, 対応分析で得られる“**全成分数**”(K)を指定したとき, 以下の関係がある. ここで, $K = \min\{m, n\} - 1$

[4-1] [階層の結合水準値の和]

= [総変動・全慣性, 固有値の和]

- 成分数を全成分数(K)より少ない成分数($K^* < K$)としたとき, 上の関係は以下のようなになる. うしろで述べる.

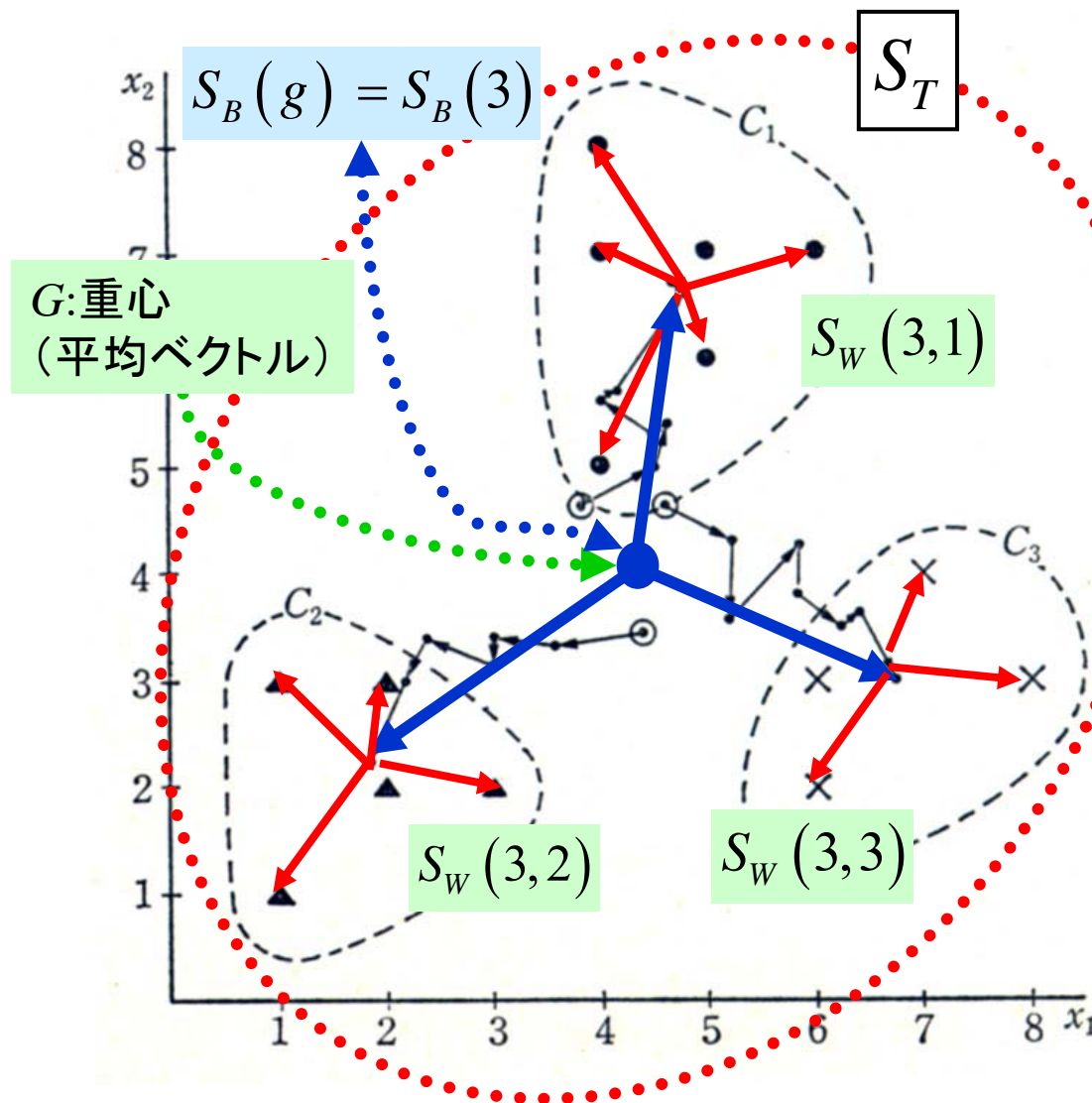
[4-2] [階層の結合水準値の和]

= [その指定した成分数 までの固有値の和]

(つづき)

- ここで, “階層の結合水準” がなにを意味するか, である.
- これは, “クラスター内変動” に対応する.
- つまり, クラスターのまとまりの程度 (等質性) を測る指標となる.
- 強固な指標とはいえないが, この変化を追うことで“クラスター数を決める” 目安とする.
- クラスター数を g (群) としたとき, その“結合の水準値の和”はその“クラスター内変動の和”に等しい.
- ここでも, 数値例でこれらの性質を調べる.

クラスター間変動とクラスター内変動(イメージ)



前に用いた図で考えよう,

クラスター数は $g=3$ と(固定)して考え方をイメージ化した.

式のおよその表現を次ページに示す.

クラスター数を変えながら学習的に目的関数を最適化する方法もある. ISODATA法とその変形. 局所的最適解となること.

クラスター間変動とクラスター内変動

- 図に合わせて各項の対応を模式的に示す.
- くりかえすが, 全変動はデータが与えられると一定に決まる値. これを2つの変動でシェアする.
- 全変動を“2つの変動”に分ける.
- 1つは, 各クラスターの重心から全重心までの平方距離, つまり“**クラスター間変動**”である.
- もう1つは, “g個”の各クラスター内の変動を加えた“**クラスター内変動の和**”である.

$$S_B(g) \Rightarrow S_B(3)$$

$$\sum_{l=1}^g S_W(g, l) \Rightarrow \sum_{l=1}^3 S_W(3, l) = S_W(3, 1) + S_W(3, 2) + S_W(3, 3)$$

$$S_T = S_B(g) + \sum_{l=1}^g S_W(g, l) \Rightarrow S_T = S_B(3) + \sum_{l=1}^3 S_W(3, l)$$

(つづき)

- たとえば, $g=3$ (群)とすると, ...

① クラスター数: g

② クラスター内変動(各クラスター内平方和): $S_W(g, l)$

↓ $g=3, l=1, 2, 3$ として,

$$S_W(3, 1), S_W(3, 2), S_W(3, 3)$$

③ これを加えるとクラスター内変動の和: $\sum_{l=1}^g S_W(g, l)$

$$\sum_{l=1}^g S_W(g, l) \Rightarrow S_W(3, 1) + S_W(3, 2) + S_W(3, 3)$$

④ クラスター間変動: $S_B(g)$ [全体の重心から各クラスター重心までの平方和]

↓

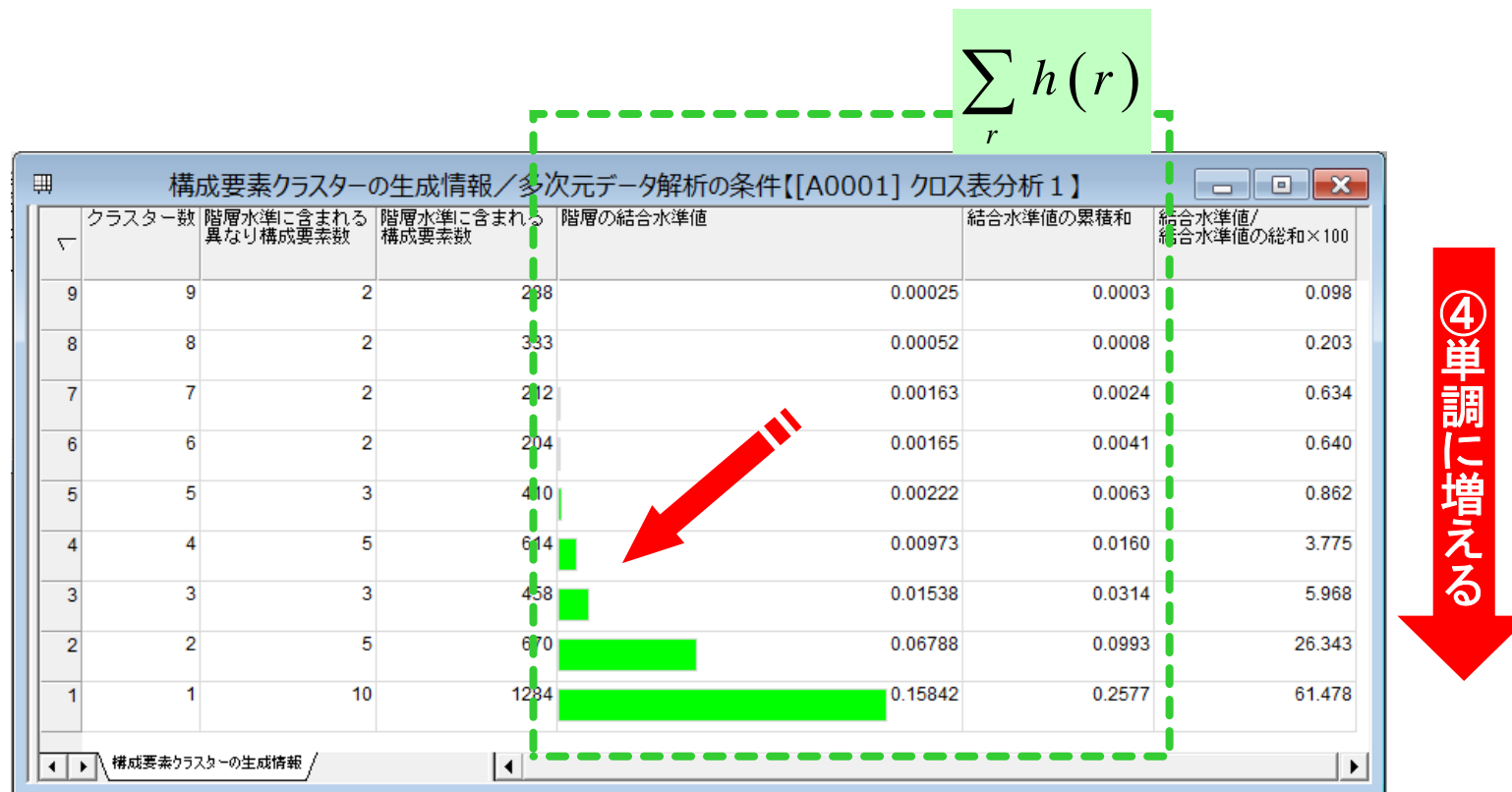
$$S_T = S_B(g) + \sum_{l=1}^g S_W(g, l)$$

どう読むか？[$g=5$ (群)として調べる]

- “クラスター数の決め方”についての明確な規則はない.
- この事情は、ヒストグラムの級数を決める問題や最適層別問題と同様、むずかしい課題である.
- クラスター化の結果は多くの場合“局所的最適解”である.
- こうしたことを知って、レストランの分類過程の履歴を観察する.
- この程度の寸法の2元データ表では、クラスター数の決め方は問題とならない.
- 実際は、かなり寸法の大きい2元データ表を扱うので“クラスター数の目安”が必要.

(つづき)

2群～3群で水準が大きく変化する. そこで5群にしてみる.



再確認: テキスト, 16ページ, 図5

$g=5$ (群) の構成 (統計量の要約)

表 11 図 9 の書き替え (説明用)

①	②	③	④	⑤	⑥	⑦		⑧	
クラスター $g = 5$ (群) $l = 1, 2, \dots, 5$	クラスター 内変動 $S_w(g, l)$	クラスター内のレストラン数とその割合		クラスター内の回答者数	クラスター重心から原点までの距離	クラスターの重心		成分スコアを使った判定	
		クラスター・サイズ (レストラン)	クラスター・サイズ 構成比	構成要素数 (回答者)	距離 (平方カイ二乗距離)	成分 スコア 1	成分 スコア 2	検定値 1	検定値 2
1	0.0005	2	0.2	333	0.1658	0.1819	-0.3643	0.61	-2.23
2	0.0000	1	0.1	125	0.0584	0.2198	0.1002	0.49	0.41
3	0.0016	2	0.2	212	0.6682	0.7667	0.2835	2.59	1.74
4	0.0016	2	0.2	204	0.0926	-0.2919	-0.0861	-0.98	-0.53
5	0.0025	3	0.3	410	0.2433	-0.4660	0.1616	-2.06	1.3
	0.0062	(10)		(1,284)					

(★★)

統計量の確認

- 「クラスターの重心」は、そのクラスターに含まれる点（レストラン）の成分スコアの平均＝そのクラスターの“重心”。
- クラスター内変動とその和（0.0062）、次ページの表の $g=5$ の欄（★★）。
- 検定の欄（⑧）はクラスター評価の指標の1つ。正規分布近似で「 ± 1.96 」が目安（有意水準5%点）。
- 形式的には、成分スコア1でこの値が大きいクラスター3と5は第1成分で説明ができる、成分スコア2で値が大きいクラスター1はこの軸で説明できる。
- このあとの、布置図で確認してみよう。

詳しく確認すると, ... (見るだけでよい)

$$S_T = S_B(g) + \sum_{l=1}^g S_W(g, l) \quad (\text{これかなり立つことをチェックしただけ})$$

表 16 総変動, クラスター間変動, クラスター内変動の関係

項目 統計量	クラスター ($g = 5$ $l = 1, 2, \dots, 5$)	⑨ 変動の 大きさ	⑩ クラスター・サイズ (クラスター内のレ ストラン数)	⑪ 構成要素数 (クラスター内の サンプル数) 表 11 の⑤に同じ	⑫ 距離 (重心からのカイニ 乗距離の二乗) 表 11 の⑥の同じ
クラスター間変動 $S_B(g)$	—	(0.2514)	(★) $S_B(5) = 0.2514$		
クラスター内変動 $S_W(g, l)$	1	0.0005	2	333	0.1658
	2	0.0000	1	125	0.0584
	3	0.0016	2	212	0.6682
	4	0.0016	2	204	0.0926
	5	0.0025	3	410	0.2433
クラスター内変動 の和 $\sum_{l=1}^g S_W(g, l)$	—	(0.0062) うしろの表 17 の④の $g = 5$	(10) (行の要素数)	(1,284) (総和)	
総変動 (全分散) S_T (固有値の和)	—	0.2577 (0.2576)	(★★★)		
変動比 [クラスター間変動/総 変動] $\eta_g = \frac{S_B(g)}{S_T} \times 100$		0.9756 (97.6%)	$\eta_g = \frac{S_B(g)}{S_T} \times 100$		
			クラスター化の程度 を測る目安とする		

5群のクラスターとその重心

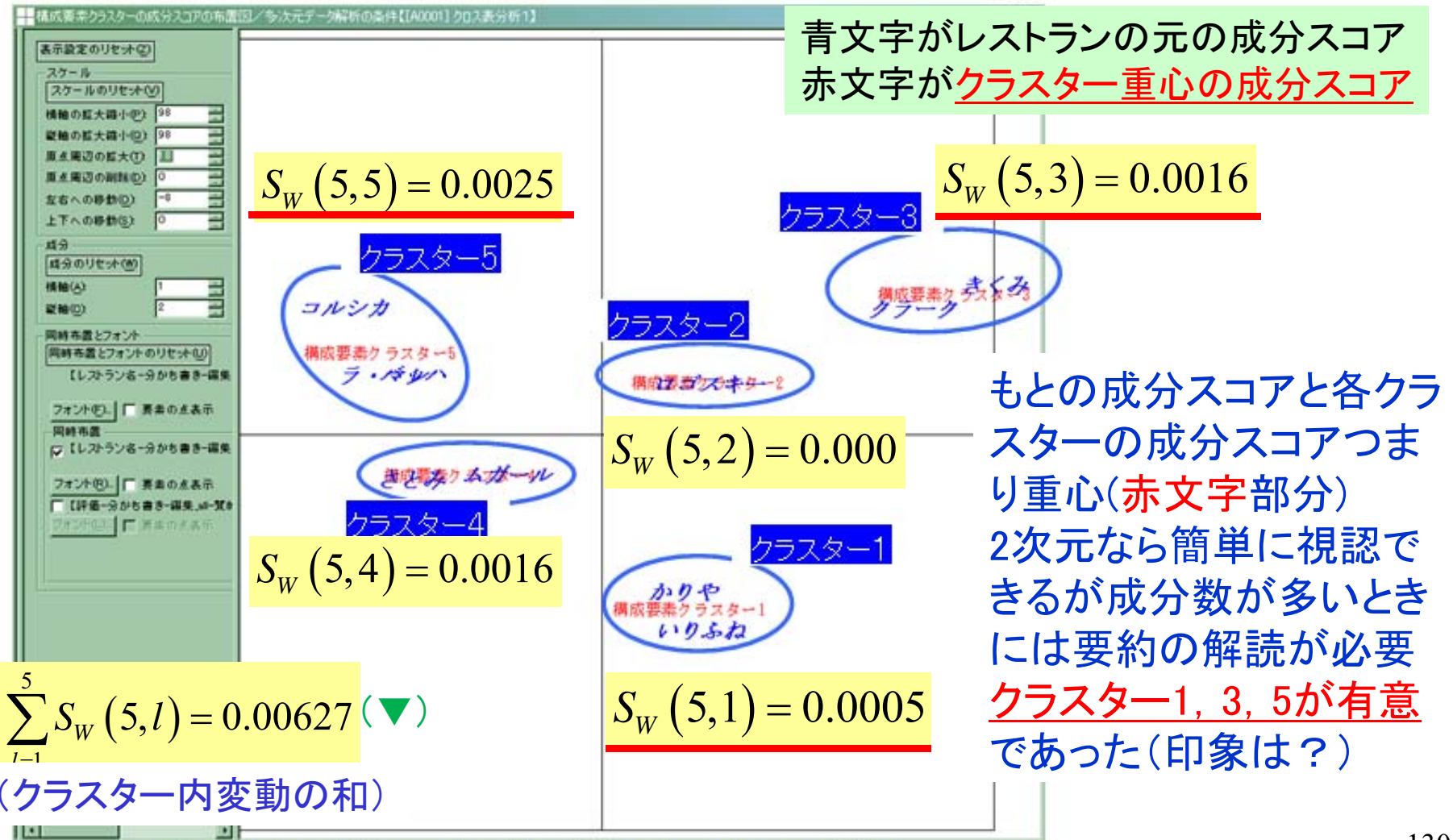


図 11 もとの成分スコアと5群の場合のクラスターの布置図

5群のクラスター構成をクロス表で再確認

表 15 5 群の場合のクラスター構成

クラスター	クラスター・サイズ (クラスター内のレストラン数)	クラスター化履歴	工夫・サービス	味	量
1	2	{いりふね, かりや}	203	60	70
2	1	{ロゴスキー}	48	35	42
3	2	{クラーク, きくみ}	69	22	121
4	2	{さとみ, ムガール}	91	90	23
5	3	{バッハ, コルシカ, ラ・マレ}	129	235	46

(10)

再確認: 表20, 34ページから

レストランを5群にクラスター化した場合の“圧縮化したクロス表”. つまり「5群×3 (評価基準)」のクロス表.

<5 群に分類後の併合を以下で追跡>

クラスター化履歴		工夫・サービス	味	量
5 群	{さとみ, ムガール}	91	90	23
	{バッハ, コルシカ, ラ・マレ}	129	235	46
	{ロゴスキー}	48	35	42
	{いりふね, かりや}	203	60	70
	{クラーク, きくみ}	69	22	121

↓

$$\chi_p^2(5) = 322.799$$

クラスター間変動

$$S_p(5) = 0.2514$$

$$0.251401 \times 1284 = 322.799$$

$$\text{クラスター内変動の和} = 0.0062$$

$$\frac{\chi_p^2(5)}{N} = 0.2514$$

(★)

(▼)

131

クラスター化に用いる成分数(K)を変えること

- ここまでは, クラスター化時に全成分数 $K=\min\{m,n\}-1$ を用いるとして議論した.
- もとの2元データ表の寸法が大きい場合が日常的である.
- こうしたとき, 成分数(K)は, 非常に大きくなる. また, 高い寄与率も望めない.
- 各成分スコアが(選択肢の)“合成変数”であることを考えると, 固有値(分散)の大きい初めのほうの成分スコアを用いてクラスター化を行う意味がある.
- つまり寄与率の変化を観察しながら, 成分数をおある K^* $< K$ とすることが必要になる.

[性質4]の後半の[4-2]の確認

[性質4]

- クラスタ化で, 対応分析で得られる“全成分数”(K)を指定したとき, 以下の関係がある.
 - [4-1] [階層の結合水準値の和]=[固有値の和]
- 成分数を全成分数(K)より少ない成分数($K^* < K$)としたとき, 上の関係は以下のようになる.
 - [4-2] [階層の結合水準値の和]
=[その指定した成分数 までの固有値の和]
- レストランの例で, 成分数を $K^*=1$ としてみよう.
- 全成分, つまり $K=2$ としたときの全変動(固有値の和)が, どう変わるか.

JMPスクリプトは成分数指定の機能はない(全成分を利用).

(つづき)

- 全成分を指定するのではなく、初めの方の何成分かを指定する ($K^* < K$).
- このときの成分数指定のルールはとくにない。 寄与率などを目安とする。
- 注意点は、指定成分数を変える(減らす)ということは、用いる分散の量(固有値の数と和)が変わる，ということ。
- つまり、元の2元データ表の総変動(全慣性)の情報のすべてを使っていない，ということ。
- これを調べておこう。

数値例で確認

- 結果を数値例で確認する.
- WordMinerの出力情報とその再編集, チェック.

構成要素クラスターの生成情報/多次元データ解析の条件【[A0003] Test_1 成分としたとき】

√	クラスター数	階層水準に含まれる 異なり構成要素数	階層水準に含まれる 構成要素数	階層の結合水準値	結合水準値の累積和	結合水準値/ 結合水準値の総和×100
9	9	2	237	0.00000	0.0000	0.000
8	8	2	280	0.00002	0.0000	0.009
7	7	3	458	0.00017	0.0002	0.087
6	6	3	383	0.00030	0.0005	0.152
5	5	4	505	0.00115	0.0016	0.584
4	4	2	212	0.00152	0.0032	0.767
3	3	5	614	0.00461	0.0078	2.331
2	2	5	670	0.03724	0.0450	18.841
1	1	10	1284	0.15265	0.1977	77.229

構成要素クラスターの生成情報

図12 クラスター生成情報(1成分のみ $K^*=1$ を指定のとき)

(つづき)

- 結合水準の和＝第1固有値(のみ)となっている.
- デンドログラムの(相対的な)高さが⑤欄となる.



表 18 図 12 の情報の要約

クラスター数 (g)	①階層水準に 含まれる 異なり構成要素数	②階層水準に 含まれる 構成要素数	③階層の 結合水準	④水準の 累積和	⑤全変動を 100 とした ときの水準の割合 (%)
9	2	237	0.00000	0.00000	0.00
8	2	280	0.00002	0.00002	0.01
7	3	458	0.00017	0.00019	0.09
6	3	383	0.00030	0.00049	0.15
5	4	505	0.00115	0.00164	0.58
4	2	212	0.00152	0.00316	0.77
3	5	614	0.00461	0.00777	2.33
2	5	670	0.03724	0.04501	18.84
1	10	1284	0.15265	0.19766 ($= \lambda_1$)	77.23 (0.15265/0.19766) *100
—	(レストランに対応)	(回答者に対応)	0.19766 ($= \lambda_1$) [結合水準の和]	—	—

第1固有値

確認とまとめ

- 対応分析法の特性, とくに“カイ二乗統計量”とその複数の変動への分解によるクラスター化.
- “カイ二乗距離”に替わり“成分スコア”の平方ユークリッド距離によるワード基準によるクラスター化.
- クラスターの成分スコア, つまりクラスター重心の求め方(加重平均), 重心からの距離(カイ二乗距離).
- 成分スコアとくに重心の成分スコアを用いたクラスター評価の検定統計量の算出とクラスターの評価(これの詳細はテキストを参照).

(つづき)

- 検定はかなりラフな指標であるが、クラスターの見当をつけるときに役立つだろう.
- “クラスター数の決め方”については、明確な決まりはない. さまざまな提案があるが、一長一短である.
- “階層の結合水準”の変化量をみる、という方式がある.
- これは、“クラスター内変動の和”の推移の変化を観察することに相当する.
- 単調に変化する値であるから、これが、急に増大したら、(平方和基準という意味での)クラスターの等質性が崩れ、変動が大きくなったとみる.

例による確認:「旅行年報2015版」から

- 初回にみた「旅行年報2015版」^(†)の「日本人の海外旅行」の項にある「旅行先別の最も楽しみにしていたこと」の要約表(頻度表)を再び用いる.
- この表からどのような情報抽出が可能か,「旅行先」の各地域と「楽しみ」にある各項目の間にどのような“関係がみられる”のだろうか,を知ること.
- この例は,(データ収集に)いくつか疑問要素があるが,これには触れない.
- 形式的に対応分析を適用,出力情報を観察する.説明を付けず,出力を順に示す.

(†)(公益財団法人)日本交通公社が毎年発行する報告書

- まず, 対応分析法を適用して得られた「同時布置図」で観察する. 情報の一部を視認しているだけ.
- 成分スコアを用いた“クラスター化”(ワード基準による分類)の結果確認する. クラスター数の見当をつける.
- 寄与度(絶対寄与度, 相対寄与度)の情報を検討する.
- 以上を, 総合的に探査, 分析してみる.
- どう客観的にデータ探査を進めるかが重要になる.
- 「机上演習」として, この例を観察してみよう.

旅行先	文化的な名所を見ること	おいしいものを食べること	自然景観を見ること	街や都市を訪れること	観光・文化施設を訪れること	スポーツやアウトドア	買い物をすること	自然の豊かさ体験	帰省・冠婚葬祭ほか	地域の文化を体験	エステ・スパ・マッサージなど	地域の祭りやイベント	目当ての祝初施設に泊まる	その他	小計
韓国	103	235	36	71	51	17	112	8	20	18	23	21	5	41	761
中国	84	38	31	39	19	8	9	6	20	17	2	4	4	11	292
台湾	158	251	79	99	71	16	32	10	18	22	7	6	4	16	789
香港・マカオ	42	53	16	49	50	15	16	6	5	7	6	3	5	9	282
シンガポール	34	30	22	57	58	8	15	5	11	10	4	6	11	6	277
インドネシア	30	11	15	8	13	29	3	12	10	9	19	1	4	8	172
マレーシア	19	17	23	29	11	10	4	13	10	9	5	1	3	10	164
タイ	78	65	24	42	9	58	20	17	9	19	15	4	6	21	387
その他東南アジア	144	37	68	47	11	44	16	13	19	25	15	1	7	18	465
オーストラリア・ニュージーランド	8	5	72	12	14	26	3	27	7	5	2	2	1	8	192
南太平洋	4	1	9	3	4	5	1	5	1	3	0	1	0	1	38
ハワイ	21	78	152	54	45	92	103	79	36	12	6	6	15	24	723
グアム・サイパン	8	25	45	14	21	123	45	28	16	8	3	2	6	7	351
アメリカ本土	18	22	80	79	70	18	30	22	45	16	0	10	3	19	432
カナダ	7	2	49	11	5	8	2	18	8	4	0	2	0	9	125
フランス	117	27	29	61	50	4	21	6	6	9	1	2	0	8	341
イギリス	39	6	13	34	21	3	13	6	5	6	0	1	0	7	154
スペイン	91	16	13	27	23	6	3	4	2	9	0	1	0	4	199
イタリア	120	17	35	58	23	7	9	6	1	6	0	0	1	6	289
ドイツ	70	7	13	48	18	0	6	6	6	8	0	9	2	10	203
その他ヨーロッパ	143	21	104	82	46	10	10	15	6	18	0	4	0	23	482
その他	89	6	80	37	10	23	7	14	3	13	2	3	0	16	303

- 各セル内の数値は、回答頻度数となっている。
- 対応分析では、この頻度データ表の行・列から同時的に眺めた割合(プロファイルという)を観察する。
- 単に、行プロファイル、列プロファイルを集計値として観察するだけではない。

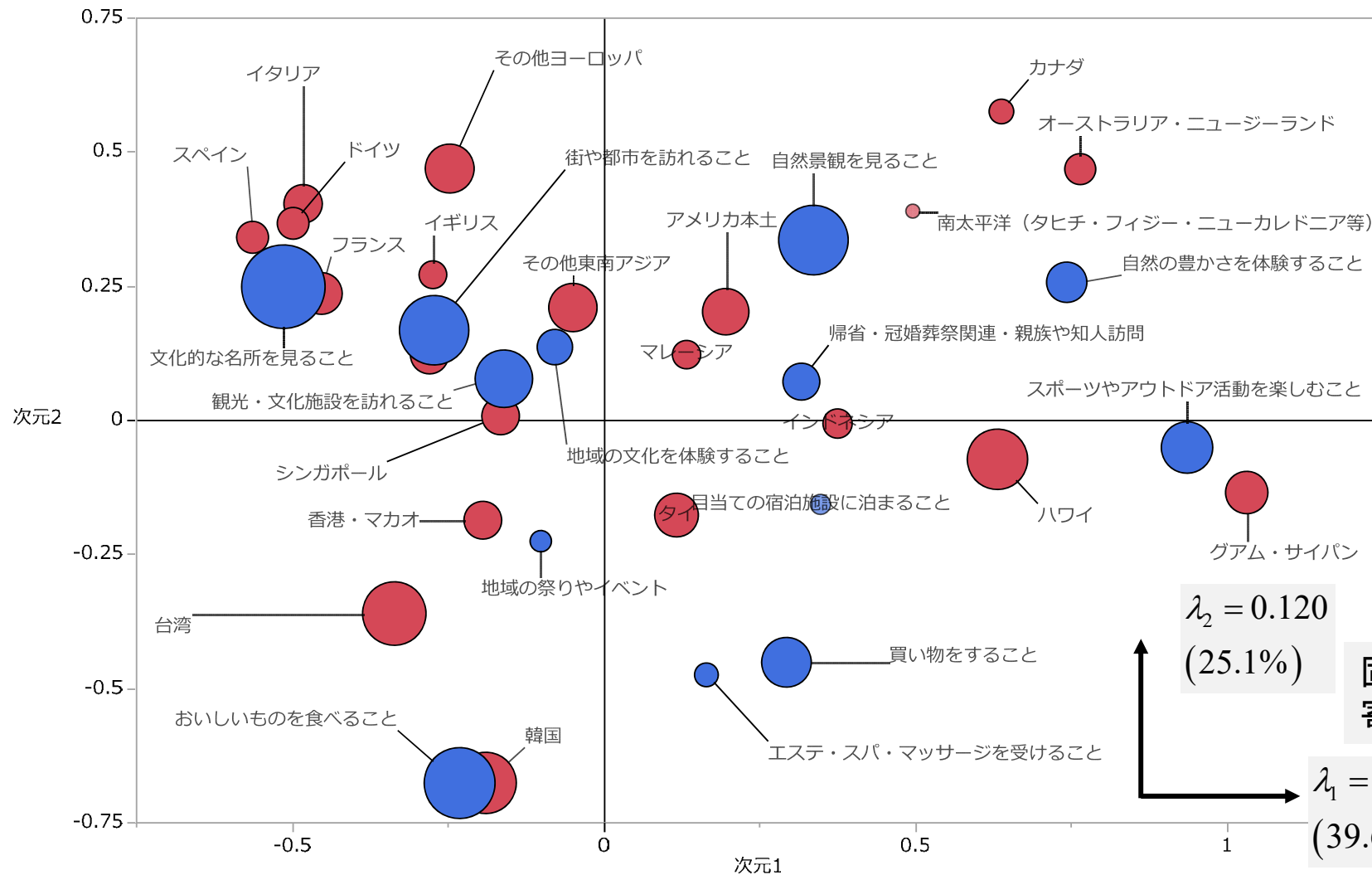
基本情報の確認

- 固有値, 特異値と寄与率, 累積寄与率を観察.
- (形式的には)はじめの3成分で約78%の情報量があると読む.
- 残りの22%程度は, 布置図の観察では見ない.
- しかし, 寄与度の評価で別の情報がみえるかもしれない.

結果						
次元	特異値	固有値	割合(%)	割合のプロット	累積(%)	累積のプロット
1	0.43447	0.18877	39.2		39.2	
2	0.34446	0.11865	24.6		63.8	
3	0.25954	0.06736	14.0		77.7	
4	0.19442	0.0378	7.8		85.6	
5	0.16191	0.02621	5.4		91.0	
6	0.1277	0.01631	3.4		94.4	
7	0.10043	0.01009	2.1		96.5	
8	0.07385	0.00545	1.1		97.6	
9	0.06836	0.00467	1.0		98.6	
10	0.05529	0.00306	0.6		99.2	
11	0.05103	0.0026	0.5		99.8	
12	0.03336	0.00111	0.2		100	

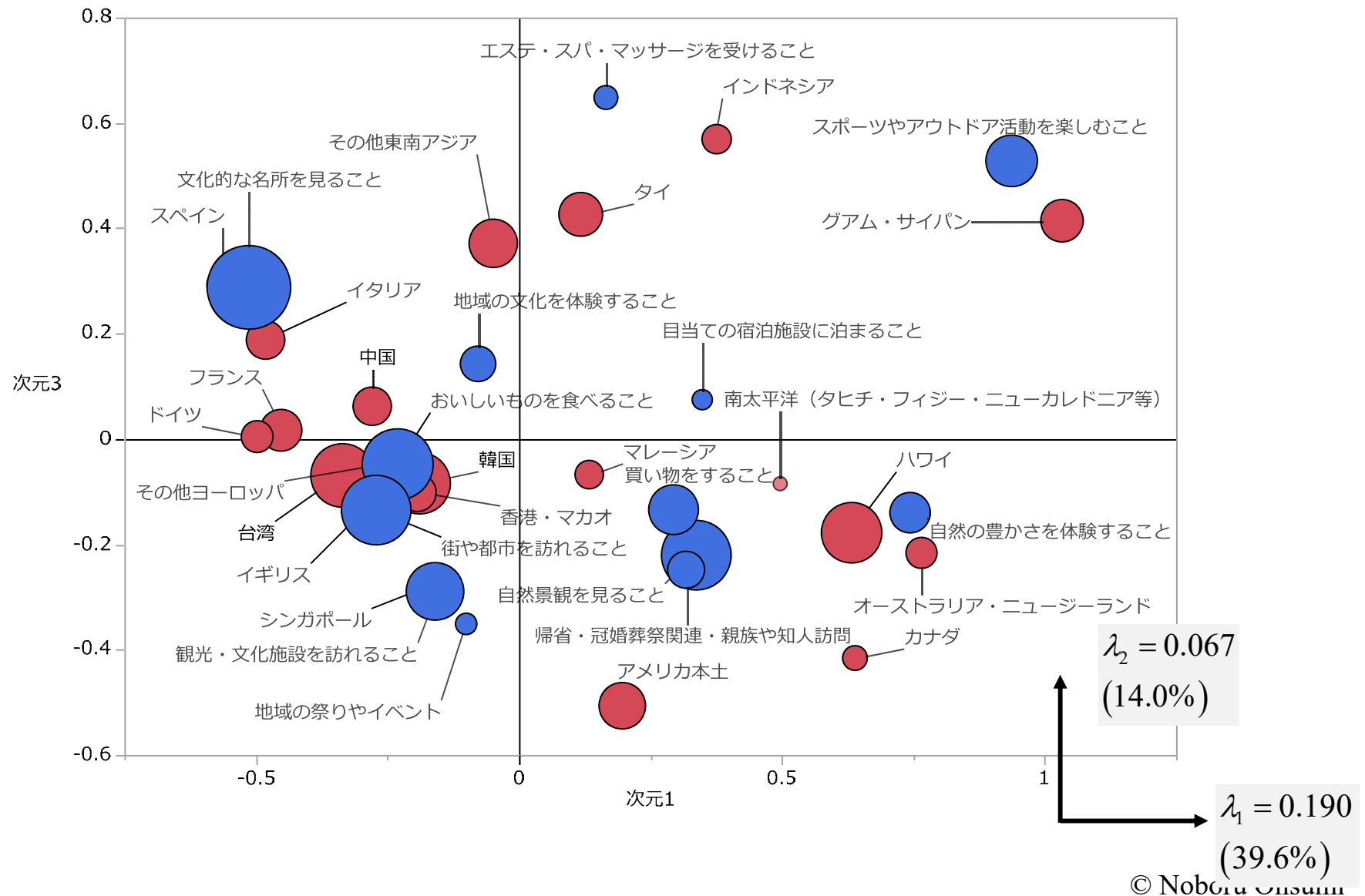
固有値の合計 = 0.482088573025273

「同時布置図」の観察(1), (1, 2)成分

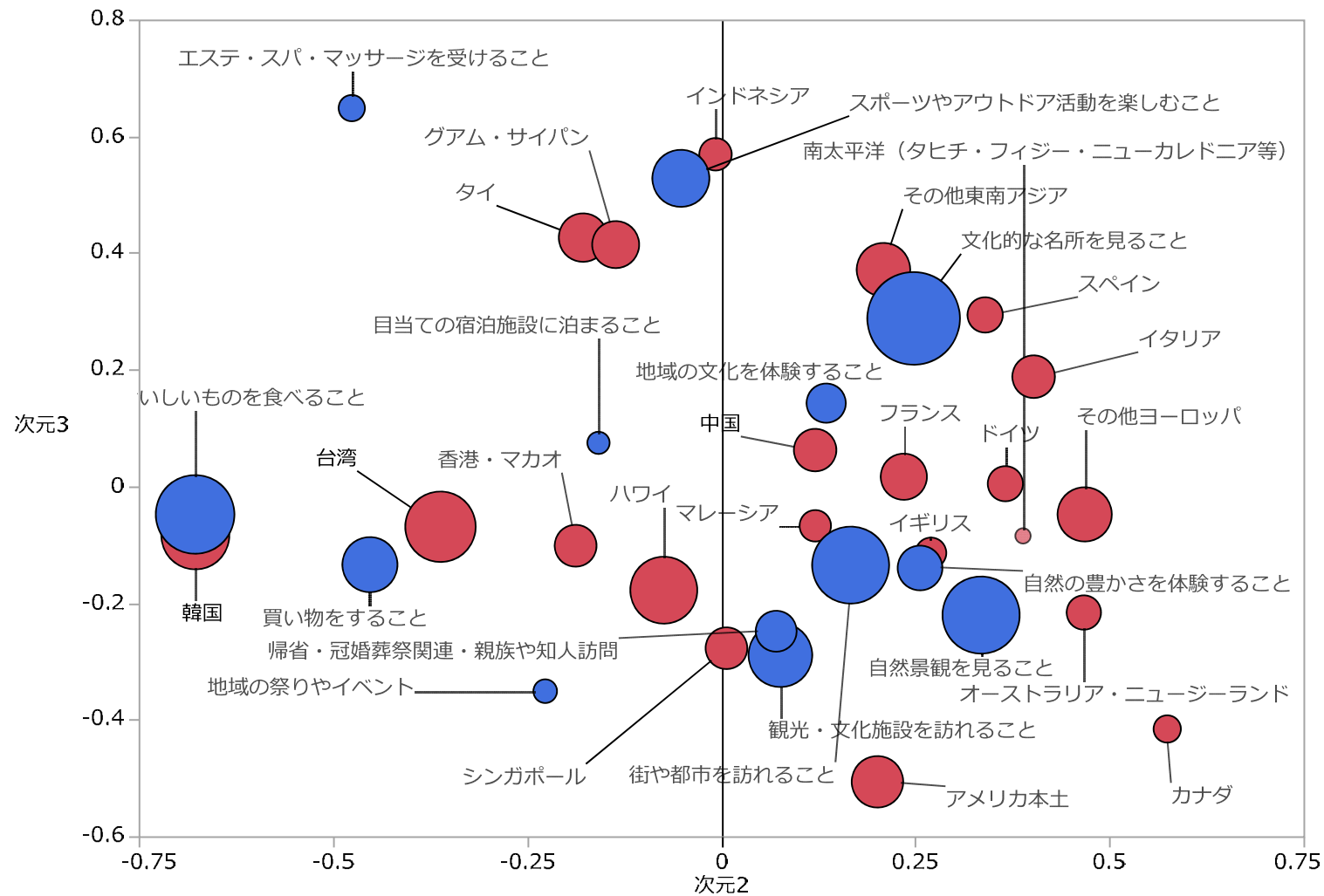


バブルの大きさは回答頻度の大きさに対応

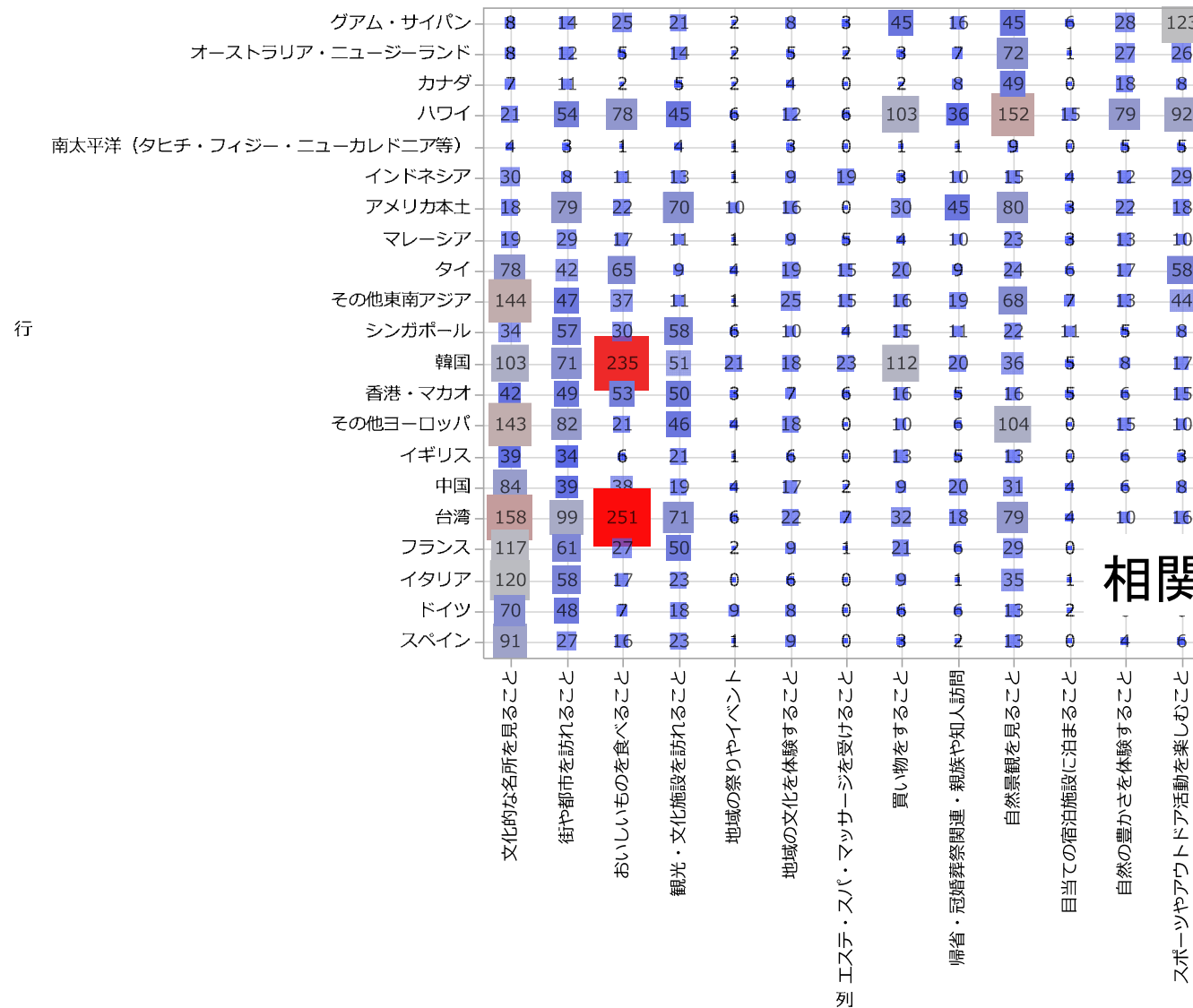
「同時布置図」の観察(1), (1, 3)成分



「同時布置図」の観察(1), (2, 3)成分



第1成分スコア之行・列の並べ替え(双対散布図)



相関係数=0.434

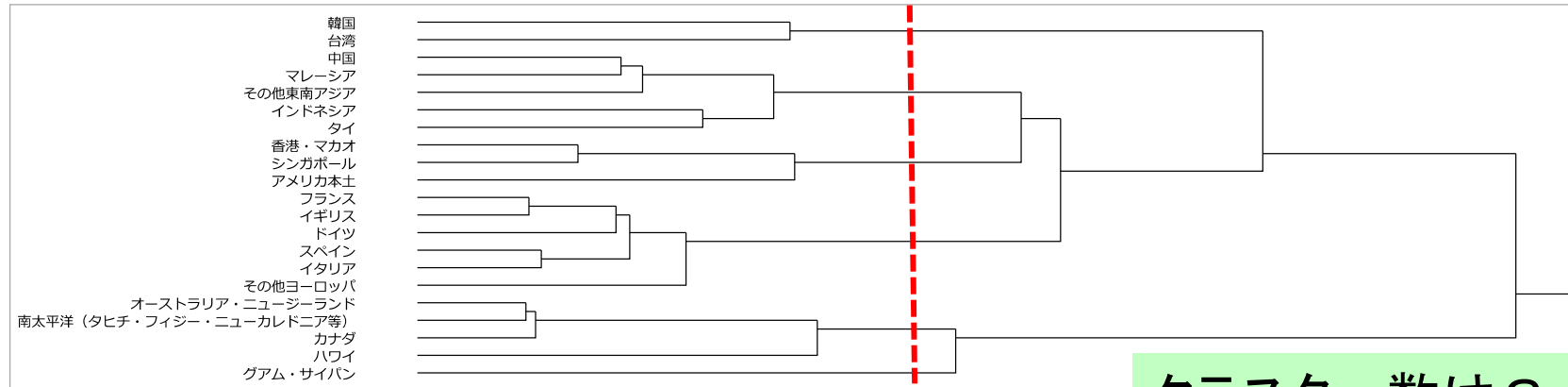
どう観察するのか？

- ここで、「旅行先の国」(行)と「旅行先での楽しみ」(列)の各要素の関係は、ある“傾向がある”ことが見える。
- 具体的には“どのように”読むのか(解釈するのか)。
- ここでは少数次元の成分スコアの平面内の情報として表示している(なぜ、そうできたのか)。
- これで十分なのか。情報はこれで尽くされているのか。
- 布置図だけではみえない情報があるのか、それはどう調べるのか。寄与度の利用は？
- データ表の情報を“視覚化”するだけでなく、さらに踏み込んで洞察するにはどうするか。



- ここから, ソフトの出力を若干加工して要約した情報をまとめて挙げる.
- 行(国)と列(旅行先での楽しみ)との分類結果
- 行と列との絶対寄与度, 相対寄与度の一覧のみ示す
- ここまでに挙げた情報と合わせて, 総合的に探査, 分析を進めてみる.
- 肝心のデータ収集環境(構成概念, 調査設計, 調査票など)の情報がないので, あくまでも推測となること.
- とりあえずおおよそのデータの特徴, 傾向の探査を机上演習としてみる.

行側(国)の分類結果

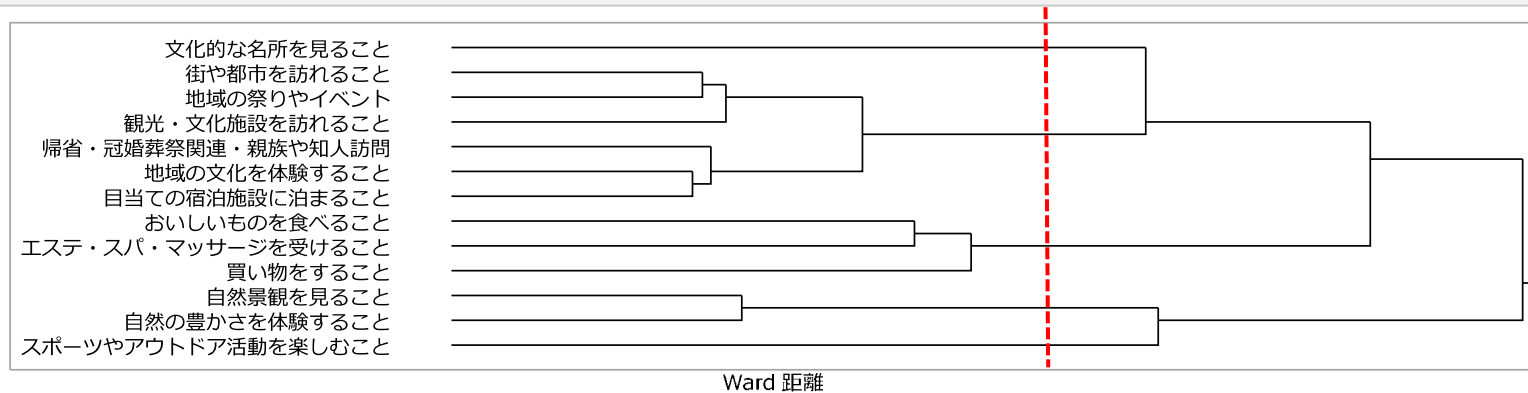


クラスター数は？

クラスター数	Ward 距離	Ward 距離の2乗	Ward 距離の2乗の累積 (クラスター内変動の和)	総変動(全慣性)に占める割合 (デンドログラムの結合高さ)	併合先の群	
20	0.03782	0.00143	0.00143	0.30	オーストラリア・ニュージーランド	南太平洋(タヒチ・フィジー・ニューカレドニア等)
19	0.03890	0.00151	0.00294	0.31	フランス	イギリス
18	0.04117	0.00170	0.00464	0.35	オーストラリア・ニュージーランド	カナダ
17	0.04317	0.00186	0.00650	0.39	スペイン	イタリア
16	0.05597	0.00313	0.00963	0.65	香港・マカオ	シンガポール
15	0.06926	0.00480	0.01443	1.00	フランス	ドイツ
14	0.07095	0.00503	0.01946	1.04	中国	マレーシア
13	0.07402	0.00548	0.02494	1.14	フランス	スペイン
12	0.07849	0.00616	0.03110	1.28	中国	その他東南アジア
11	0.09367	0.00877	0.03987	1.82	フランス	その他ヨーロッパ
10	0.09947	0.00989	0.04976	2.05	インドネシア	タイ
9	0.12423	0.01543	0.06519	3.20	中国	インドネシア
8	0.12988	0.01687	0.08206	3.50	韓国	台湾
7	0.13153	0.01730	0.09936	3.59	香港・マカオ	アメリカ本土
6	0.13942	0.01944	0.11880	4.03	オーストラリア・ニュージーランド	ハワイ
5	0.18767	0.03522	0.15402	7.31	オーストラリア・ニュージーランド	グアム・サイパン
4	0.21050	0.04431	0.19833	9.19	中国	香港・マカオ
3	0.22422	0.05027	0.24860	10.43	中国	フランス
2	0.29469	0.08684	0.33544	18.01	韓国	中国
1	0.38292	0.14663	0.48207	30.42	韓国	オーストラリア・ニュージーランド

列側（旅行先での楽しみ）の分類

樹形図



クラスター数	Ward 距離	Ward 距離の2乗	Ward 距離の2乗の累積(クラスター内変動の和)	総変動(全慣性)に占める割合(デンドログラムの結合高さ)	併合先の群	併合された群
12	0.08396	0.00705	0.00705	1.46	地域の文化を体験すること	目当ての宿泊施設に泊まること
11	0.08743	0.00764	0.01469	1.58	街や都市を訪れること	地域の祭りやイベント
10	0.09037	0.00817	0.02286	1.69	帰省・冠婚葬祭関連・親族や知人訪問	地域の文化を体験すること
9	0.09559	0.00914	0.03200	1.90	街や都市を訪れること	観光・文化施設を訪れること
8	0.10112	0.01022	0.04222	2.12	自然景観を見ること	自然の豊かさを体験すること
7	0.14317	0.02050	0.06272	4.25	街や都市を訪れること	帰省・冠婚葬祭関連・親族や知人訪問
6	0.16126	0.02600	0.08872	5.39	おいしいものを食べること	エステ・スパ・マッサージを受けること
5	0.18097	0.03275	0.12147	6.79	おいしいものを食べること	買い物をする
4	0.24178	0.05846	0.17993	12.13	文化的な名所を見ること	街や都市を訪れること
3	0.24616	0.06059	0.24052	12.57	自然景観を見ること	スポーツやアウトドア活動を楽しむこと
2	0.31997	0.10238	0.34290	21.24	文化的な名所を見ること	おいしいものを食べること
1	0.37307	0.13918	0.48208	28.87	文化的な名所を見ること	自然景観を見ること

ここもクラスター数は？

行(国)の絶対寄与度(1)

カテゴリ名	周辺度数	周辺割合(%)	次元1	次元2	次元3	次元4	次元5	次元6
韓国	720	10.5	1.98	40.59	1.11	0.95	0.31	19.43
中国	281	4.1	1.69	0.51	0.23	0.13	1.36	0.92
台湾	773	11.3	6.72	12.41	0.81	12.05	0.06	31.77
香港・マカオ	273	4	0.79	1.18	0.62	7.28	0.31	9.32
シンガポール	271	4	0.57	0.00	4.52	21.09	2.73	3.53
インドネシア	164	2.4	1.80	0.00	11.50	1.69	38.04	1.07
マレーシア	154	2.2	0.21	0.28	0.15	0.01	7.68	0.00
タイ	366	5.3	0.39	1.42	14.40	0.00	1.66	0.13
その他東南アジア	447	6.5	0.08	2.40	13.34	2.53	2.95	3.26
オーストラリア・ニュージーランド	184	2.7	8.34	4.94	1.87	10.43	1.13	4.01
南太平洋(タヒチ・フィジー・ニューカレドニア等)	37	0.5	0.71	0.69	0.06	0.18	0.03	0.30
ハワイ	699	10.2	21.65	0.47	4.81	1.59	5.18	9.88
グアム・サイパン	344	5	28.38	0.78	12.78	8.56	16.67	7.32
アメリカ本土	413	6	1.24	2.06	22.97	9.91	3.37	0.12
カナダ	116	1.7	3.67	4.71	4.35	11.60	1.48	0.01
フランス	333	4.9	5.26	2.27	0.02	2.16	4.99	0.22
イギリス	147	2.1	0.85	1.32	0.42	3.17	1.83	1.61
スペイン	195	2.8	4.77	2.77	3.64	0.01	2.45	0.11
イタリア	283	4.1	5.07	5.64	2.16	0.27	6.20	0.17
ドイツ	193	2.8	3.69	3.18	0.00	2.59	0.22	6.43
その他ヨーロッパ	459	6.7	2.15	12.38	0.23	3.80	1.32	0.35
			100.00	100.00	100.00	100.00	100.00	100.00

絶対寄与度の傾向は？ マーカーでチェックしてみよう。
それを、同時布置図と合わせて観察する。

行(国)の絶対寄与度(2)

カテゴリ名	周辺度数	周辺割合(%)	次元7	次元8	次元9	次元10	次元11	次元12
韓国	720	10.5	0.53	0.13	11.26	2.88	0.03	0.83
中国	281	4.1	22.50	0.75	6.14	0.06	11.19	0.20
台湾	773	11.3	2.25	0.09	1.00	1.90	0.00	3.19
香港・マカオ	273	4	8.95	0.18	0.00	0.25	0.18	0.05
シンガポール	271	4	4.39	3.89	8.70	21.66	4.33	0.57
インドネシア	164	2.4	10.80	5.58	3.52	1.96	0.38	6.07
マレーシア	154	2.2	0.49	5.86	5.77	11.82	8.11	3.78
タイ	366	5.3	0.97	13.76	0.00	1.73	4.19	4.03
その他東南アジア	447	6.5	2.99	4.79	5.88	13.31	0.54	1.22
オーストラリア・ニュージーランド	184	2.7	4.20	0.80	6.58	0.50	1.62	0.45
南太平洋(タヒチ・フィジー・ニューカレドニア等)	37	0.5	0.00	2.52	4.56	1.66	9.05	25.97
ハワイ	699	10.2	5.75	0.06	23.30	1.24	1.95	0.54
グアム・サイパン	344	5	5.19	0.29	3.81	0.92	0.13	0.46
アメリカ本土	413	6	18.96	10.23	4.08	0.64	2.98	1.28
カナダ	116	1.7	0.09	1.08	1.73	0.34	0.31	1.17
フランス	333	4.9	4.72	7.78	0.68	5.06	0.53	0.08
イギリス	147	2.1	0.40	0.14	0.01	12.53	5.97	11.79
スペイン	195	2.8	0.10	1.93	0.28	5.59	29.80	0.61
イタリア	283	4.1	2.06	0.07	2.33	0.29	8.56	16.10
ドイツ	193	2.8	2.76	39.72	6.30	0.04	1.81	13.05
その他ヨーロッパ	459	6.7	1.91	0.35	4.08	15.64	8.35	8.53
			100.00	100.00	100.00	100.00	100.00	100.00

列(旅行先での楽しみ)の絶対寄与度

カテゴリ名	周辺度数	周辺割合(%)	次元1	次元2	次元3	次元4	次元5	次元6
文化的な名所を見ること	1338	19.5	27.25	10.12	23.99	2.23	3.79	3.06
おいしいものを食べること	964	14.1	3.95	54.39	0.49	13.00	0.16	11.24
自然景観を見ること	928	13.5	8.17	12.80	9.80	29.48	0.23	0.67
街や都市を訪れること	924	13.5	5.26	3.17	3.66	8.73	0.02	0.19
観光・文化施設を訪れること	633	9.2	1.24	0.45	11.50	26.31	0.03	14.77
スポーツやアウトドア活動を楽しむこと	507	7.4	34.43	0.17	30.59	5.38	4.84	10.77
買い物をすること	473	6.9	3.17	11.89	1.85	2.19	21.70	39.89
自然の豊かさを体験すること	312	4.6	13.36	2.52	1.33	3.89	0.27	1.35
帰省・冠婚葬祭関連・親族や知人訪問	261	3.8	2.04	0.16	3.49	2.57	10.17	3.95
地域の文化を体験すること	240	3.5	0.11	0.54	1.06	0.04	6.87	0.18
エステ・スパ・マッサージを受けること	108	1.6	0.23	3.00	9.83	0.25	48.19	6.68
地域の祭りやイベント	87	1.3	0.07	0.55	2.33	1.22	0.46	7.24
目当ての宿泊施設に泊まること	77	1.1	0.72	0.24	0.09	4.69	3.28	0.00
			100.00	100.00	100.00	100.00	100.00	100.00
カテゴリ名	周辺度数	周辺割合(%)	次元7	次元8	次元9	次元10	次元11	次元12
文化的な名所を見ること	1338	19.5	0.01	1.26	0.14	0.06	5.70	2.85
おいしいものを食べること	964	14.1	1.03	0.39	0.24	0.69	0.10	0.25
自然景観を見ること	928	13.5	0.75	2.93	0.48	18.80	2.33	0.01
街や都市を訪れること	924	13.5	0.43	16.50	0.72	1.77	45.95	0.11
観光・文化施設を訪れること	633	9.2	12.14	10.28	3.42	0.10	10.11	0.41
スポーツやアウトドア活動を楽しむこと	507	7.4	2.42	0.12	2.47	0.36	0.69	0.37
買い物をすること	473	6.9	2.58	4.61	0.56	0.37	0.84	3.44
自然の豊かさを体験すること	312	4.6	6.75	12.72	1.17	42.70	9.22	0.16
帰省・冠婚葬祭関連・親族や知人訪問	261	3.8	42.06	14.85	2.18	3.63	0.55	10.54
地域の文化を体験すること	240	3.5	8.56	0.53	0.01	0.00	1.79	76.81
エステ・スパ・マッサージを受けること	108	1.6	18.31	2.00	3.52	0.01	5.84	0.55
地域の祭りやイベント	87	1.3	3.54	25.29	35.19	9.63	9.97	3.26
目当ての宿泊施設に泊まること	77	1.1	1.41	8.52	49.89	21.87	6.91	1.23
			100.00	100.00	100.00	100.00	100.00	100.00

行(国)の相対寄与度(平方相関)

カテゴリ名	周辺度数	周辺割合(%)	次元1	次元2	次元3	次元4	次元5	次元6	6成分まで
韓国	720	10.5	6.56	84.59	1.31	0.63	0.14	5.57	98.80
中国	281	4.1	43.07	8.14	2.13	0.66	4.84	2.04	60.88
台湾	773	11.3	33.30	38.68	1.43	11.97	0.04	13.61	99.02
香港・マカオ	273	4	17.32	16.30	4.83	32.08	0.95	17.73	89.21
シンガポール	271	4	7.03	0.01	20.01	52.35	4.70	3.78	87.90
インドネシア	164	2.4	14.35	0.01	32.81	2.70	42.23	0.74	92.82
マレーシア	154	2.2	9.80	8.05	2.52	0.12	49.00	0.01	69.49
タイ	366	5.3	5.43	12.36	71.11	0.01	3.19	0.16	92.26
その他東南アジア	447	6.5	0.99	18.38	57.88	6.16	4.98	3.43	91.81
オーストラリア・ニュージーランド	184	2.7	55.07	20.50	4.40	13.78	1.03	2.29	97.07
南太平洋(タヒチ・フィジー・ニューカレドニア等)	37	0.5	41.17	25.20	1.21	2.05	0.26	1.53	71.42
ハワイ	699	10.2	81.73	1.11	6.48	1.21	2.72	3.22	96.47
グアム・サイパン	344	5	73.73	1.27	11.85	4.45	6.01	1.64	98.96
アメリカ本土	413	6	8.43	8.84	55.91	13.54	3.19	0.07	89.98
カナダ	116	1.7	33.93	27.41	14.37	21.50	1.90	0.01	99.12
フランス	333	4.9	62.47	16.93	0.08	5.13	8.23	0.23	93.07
イギリス	147	2.1	26.22	25.66	4.59	19.62	7.86	4.31	88.26
スペイン	195	2.8	54.60	19.95	14.87	0.02	3.90	0.11	93.44
イタリア	283	4.1	47.39	33.11	7.21	0.50	8.04	0.14	96.38
ドイツ	193	2.8	44.22	23.95	0.00	6.21	0.37	6.65	81.40
その他ヨーロッパ	459	6.7	18.51	66.95	0.71	6.55	1.57	0.26	94.56

相対寄与度の傾向は？ ここもチェックを入れてみよう。
また、同時布置図と合わせて観察する。

列(旅行先での楽しみ)の相対寄与度(平方相関)

カテゴリ名	周辺度数	周辺割合(%)	次元1	次元2	次元3	次元4	次元5	次元6	6次元まで
文化的な名所を見ること	1338	19.5	62.57	14.61	19.66	1.02	1.21	0.61	99.69
おいしいものを食べること	964	14.1	9.42	81.40	0.41	6.20	0.05	2.31	99.79
自然景観を見ること	928	13.5	31.22	30.73	13.36	22.55	0.12	0.22	98.19
街や都市を訪れること	924	13.5	45.70	17.30	11.35	15.21	0.03	0.15	89.73
観光・文化施設を訪れること	633	9.2	9.30	2.14	30.72	39.46	0.03	9.56	91.21
スポーツやアウトドア活動を楽しむこと	507	7.4	71.22	0.22	22.58	2.23	1.39	1.92	99.56
買い物をすること	473	6.9	17.09	40.35	3.57	2.37	16.27	18.60	98.26
自然の豊かさを体験すること	312	4.6	74.51	8.84	2.64	4.35	0.21	0.65	91.20
帰省・冠婚葬祭関連・親族や知人訪問	261	3.8	23.98	1.20	14.63	6.04	16.57	4.01	66.43
地域の文化を体験すること	240	3.5	4.04	12.34	13.66	0.28	34.62	0.56	65.51
エステ・スパ・マッサージを受けること	108	1.6	1.62	13.34	24.79	0.35	47.29	4.08	91.47
地域の祭りやイベント	87	1.3	1.57	8.06	19.40	5.71	1.48	14.61	50.84
目当ての宿泊施設に泊まること	77	1.1	16.79	3.44	0.76	21.78	10.55	0.01	53.33

JMRAマーケティング・リサーチ講座

質的データのマイニング のための対応分析法

スライド資料[その5]

大隅 昇

ohsumi@ss.ij4u.or.jp

<http://wordminer.org/>

Copyright by Noboru Ohsumi

再確認: 対応分析法で扱う2元データ表

- ① “2元(two-way)の行列”形式となっていること
- ② 各要素(セル)内の数値は“非負の値”であること
- ③ 行あるいは列の“比率のパターン”, つまり“プロフィール”を考える意味があるような場合
- ④ あるいはそれに相当する場面を想定できる行列形式のデータ表

- この条件を満たすデータ表は身近に沢山みられる.
- まず, 分析したいデータ表が要件を満たすかを確認.
- とくにいくつかの形式のデータ表間の数理的な関係が調べられている. それを知って用いる.

復習: データ表の関係をすること(重要)

- データ表(=行列)に“便宜的に名前”を付けて整理する.
- 多変量構造のデータ行列(2元)[X表]
- “2元クロス表”[F表]. これが基本となる.
- 質問文の選択肢のコーディングで得たデータ表[C表]
- アイテム・カテゴリー型行列, インジケータ行列, 完備排反型行列(以下で, “インジケータ行列”とする)[A表]
- “多重クロス表”(バート表, バート行列)[B表]

バート表とは, 提唱者のC. Burtの名前を付けたもの.
英国の応用心理学者.
多元クロス表(multi-way)と多重クロス表(multiple)の違い

[その5]で述べること

- “多重対応分析”(MCA)について, 簡単に説明する.
- 多重クロス表(バート表, バート行列)も, 2元データ表の1つの形であることはすでに述べた.
- この多重クロス表の対応分析を行うことが, もっとも一般的な“多重対応分析”である. “変形”がいろいろある.
- すでにみた2元データ表間の関連を, 再度確認する.
- まず, アイテム・カテゴリー型または“インジケータ行列”(A表)の対応分析を調べる.
- つぎに, A表から作られる“多重クロス表”(B表)の対応分析を調べる.

(つづき)

- 2項目: $[X\text{表}] \Rightarrow [F\text{表}] (2\text{元クロス表}) \Leftrightarrow [A\text{表}] \Rightarrow [B\text{表}]$
- 項目数が2項目の場合は, 特別な関係があること.
- 3項目以上: $[X\text{表}] \Rightarrow [C\text{表}] \Rightarrow [A\text{表}] \Rightarrow [B\text{表}]$
- これらの間には, 共通した性質があること(ある意味で同等となること).
- A表, B表からえた固有値, 寄与率の関係を調べること.
- データ表の見方を変えてアプローチすることで, 別の構造探査が可能となること.

以下では, 多重対応分析を「MCA」とすることがある.
同じく, 対応分析法を「CA」とすることがある.

(つづき)

- 社会調査データを例とすると、調査で得た多数の質問項目と選択肢の関係の吟味に用いる。
- こうした“回答者の(意識の)ゆらぎ”の入ったデータの構造探査だけでなく、別の変数、とくに真値ないしはそれに近い情報を加えて、多重対応分析を行うことの効用があること。
- たとえば、**人口統計学的変数**(属性ほか)や“**外部情報源**”を加えることで、分析の信頼性を上げること。
- 大がかりな例は、調査背景の説明に時間を要するので、加工した実際データやトイ・データで分析を試みること。
- うしろで簡単な例をみる。時間の許す範囲で説明。

外部情報源(external information)の活用
公的調査では必須の操作。

★参考：外部情報源 (external information) とは

- “客観情報” となり得る別情報, これを併用する効用.
- 調査などで, 実査で手にした回答データだけでなく, 二次情報, 副次情報として, 他の情報源を用いること.

例1: 卑近な例としては, 調査で対象者の実年齢報告や年間世帯所得などには, 偏りがあることはよく知られたこと. これを, 住民基本台帳や選挙人名簿情報, 税務申告などで補正・調整(補定). 公的調査など. マイナンバーもそうした情報となるだろう. ただし商用調査では困難.

例2: 携帯電話のキャリアに関する質問(どこと契約)で, 回答データだけでなく, 携帯電話の販売状況のデータを併用する.

(つづき)

例3: 病院の通院状況や診察体験を問う調査で、ある期間の通院回数を聴取したとする。こうした回答は“記憶”に依存するので“偏り”が入る。そこで、通院記録から、その人が実際に通院した回数を調べ対比する。

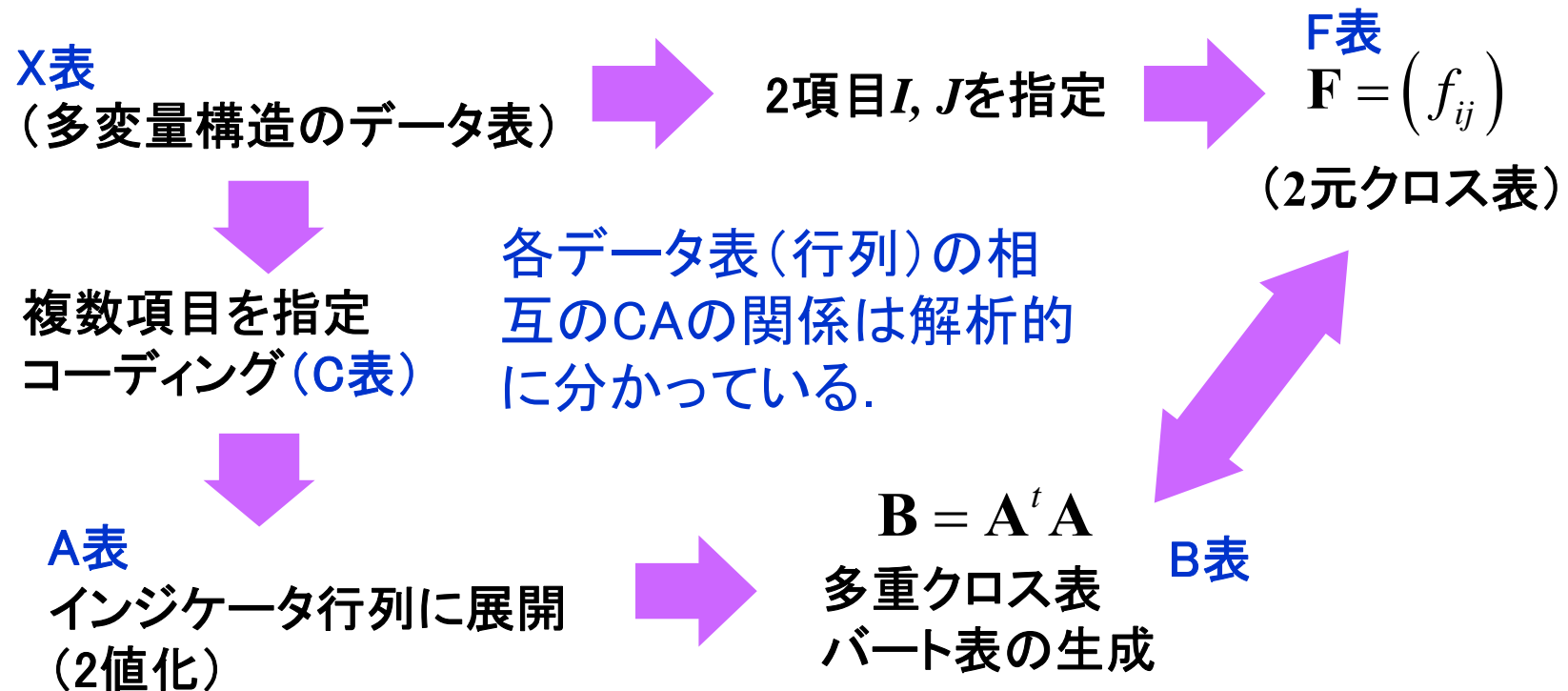
(*) 無記名記入のときにはどうするかがある。

例4: 旅行調査で、実際に旅行した回数を問う(ほぼ偏りあり)。日記形式で記録する、海外旅行ならパスポート確認などで調整。

例5: 危険行動・違法ドラッグの使用などを問うと、調査方式によっては正しい回答が得られない。調査方式を変えるだけでなく、犯罪履歴をしらべ、自己申告と比較する(米国の例/全国犯罪被害調査:NCVS)。社会的望ましさによる偏り。

再確認: 要約: おもな2元データ表の相互の関連

- 「第Ⅰ部」の最後に「付表」として整理してある中の一部.
- ここまでに述べたデータ表の関係を模式図にする.
- 「第Ⅱ部」にも説明がある.



★参考：再コーディング，コード変換の応用

- 1つの発展型として再コード化 (recoding)，コード変換を用いて，さまざまな(変形)データ表を作り，これにMCAを適用する方法が提案されている.
- 再コーディングの別の例として，“量的データ”をカテゴリー化する(区分化)簡単な例を示した.
- 「血圧」「世帯所得」「Web調査回答所要時間」など.
- 「年齢区分」の作り方なども要注意の項目である.
- 尺度をコーディングで順序尺度化するような場合.
- これは、実は難しい課題の1つであるとも指摘した.

(つづき)

- 順序尺度データの“**反対変数・逆変数**”あるいは“**重複コード化**”を行って、MCAを行う方法もある.
- とくに、因子分析を行う前に、順序尺度の向きを調べるために、この“**反対変数**”を使ったMCAが有効なことがある.

例: SD法の選択肢の尺度の向きの確認

例: 反対変数の変換操作のヒント

$$[\text{逆コード}] = [\text{最大選択肢} + \text{最小選択肢}] - [\text{当該選択肢}]$$

- 1.非常に満足
- 2.満足
- 3.ふつう
- 4.あまり満足ではない
- 5.まったく満足ではない

ここで選択肢「4」を選ぶとその反対コードは？

反対変数・逆変数(counter variable), 重複コード化(double coding)

(つづき)

- これを“2値データ表”に適用することもある. 結果はどうなるのだろう. 演習で試みるとよい.
- たとえば, テキスト(第 I 部)に示した, また演習用データとした「清涼飲料水データ」に適用すると, どうなるか.
- “順位データ”, “順位評価”に変換し, これにCA・MCAを適用するとき.

量的データの尺度化, 順位データ(ranking data)
ここらは、言葉としてだけ述べておく. 関連書を参照.

多重対応分析のおもな特性を例で確認

◎“2項目”の多重対応分析と元の2元データ表

- 再び「レストラン評価」データを用いる.
- 2元クロス表のCAと2項目のMCAは, 同等結果となる.
- 元のデータ表で, 行数(回答者数, 対象者数)が非常に多いとき, ある対策(追加処理)で, 計算効率をはかる.
- 2項目の場合, CA, MCAでえられる固有値, 寄与率の関係を調べること.

(つづき)

◎一般に多数項目の多重対応分析の特性

- 扱うデータ表の関係を再確認する.
- 用いる多数項目間の関連性の評価, 観察の方法.
- 項目の選択肢に付与の成分スコアの特徴.
- 固有値, 寄与率, 寄与度などの特性と解釈.
- おもに事例データで検証する. 例として以下を用いる.
- 例:「環境意識調査」データを再び確認
- 例:ある自治体の「市民意識調査」

★追加処理について(言葉のみ)

- “追加処理”とは、データの一部を除外して、再配置する処理のこと.
- 行, 列のどちらにも適用できる. たとえば, ある質問項目を一時除去と再配置, サンプルを一時除去, 再配置など.
- 計算に用いるデータを“アクティヴ要素”, 追加処理とするデータを“追加要素”という.
- バート表にインジケータ行列を追加処理することで回答者ベースの成分スコアが得られる(後述).

(*)具体的な適用場面については, テキスト, 第 I 部, 46ページあたりを参照.

追加処理(supplementary treatment), アクティヴ要素(active element), 追加要素(supplementary element)

2項目の場合の各データ表の関連(再確認)

X表(多変量構造のデータ表)

項目 回答者	<i>I</i> (レストラン)	<i>J</i> (評価基準)
1	バッハ	味
2	ムガール	量
3	さとみ	量
4	ラ・マレ	工夫・サービス
5	きくみ	味
⋮	⋮	⋮
⋮	⋮	⋮
<i>N</i>	いりふね	量

N=1,284 (回答者数)

C表(コーディング)

$$\mathbf{F} = (f_{ij})_{m \times n}$$

(2元クロス表; F表)

A表(インジケータ行列に展開) $\mathbf{A}_{N \times n^*}$

表 28 (回答者) × (アイテム・カテゴリー) のデータ表, インジケータ行列 (A 表)

項目 回答者	<i>I</i>										<i>J</i>		
	1	2	3	4	...	9	10	1	2	3	味	量	工
1	0	1	0	0	...	0	0	1	0	0			
2	0	0	1	0	...	0	0	0	1	0			
3	1	0	0	0	...	0	0	0	1	0			
4	0	0	0	0	...	1	0	0	0	1			
5	⋮	⋮	⋮	⋮	...	⋮	⋮	⋮	⋮	⋮			
⋮	⋮	⋮	⋮	⋮	...	⋮	⋮	⋮	⋮	⋮			
1,284	0	0	0	1	...	1	1	0	1	0			

$$\mathbf{B} = \mathbf{A}^t \mathbf{A}_{n^* \times n^*}$$

(多重クロス表・バート表; B表)

2項目*I*と*J*のインジケータ行列(A表)

- いままでと若干, 記号の使い方を変える.
- 多項目のMCAを一般化表記するときの都合. またテキストの記法に合わせる.
- つねに, 扱っているデータ表の“寸法”に注意すること.

①2つの項目, *I*と*J*とその選択肢

$$I = \{1, 2, \dots, n_i\}, J = J = \{1, 2, \dots, n_j\}$$

[前の記法で, $m = n_i, n = n_j$ と対応する]

②インジケータ行列を作る

$$\mathbf{A}_{N \times n^*} = \begin{bmatrix} \mathbf{A}_i & \mathbf{A}_j \\ N \times n_i & N \times n_j \end{bmatrix} \quad \left(\text{ここで } n^* = n_i + n_j \right)$$

バート表(B表)とクロス表(F表)

- バート表を作り, 2項目(項目数を $M=2$)のクロス表がどう書けるかをみる.

③バート表を行列表記する

$$\mathbf{B}_{n^* \times n^*} = \mathbf{A}_{n^* \times N}^t \mathbf{A}_{N \times n^*} = \begin{pmatrix} \mathbf{A}_i^t \mathbf{A}_i & \mathbf{A}_i^t \mathbf{A}_j \\ \mathbf{A}_j^t \mathbf{A}_i & \mathbf{A}_j^t \mathbf{A}_j \end{pmatrix} \quad \left(\text{ここで } n^* = n_i + n_j \right)$$

④2項目のクロス表を行列表記

$$\mathbf{F}_{n_i \times n_j} = \mathbf{A}_i^t \mathbf{A}_j \quad (\text{項目 } I \times J \text{ のクロス表})$$

$$\mathbf{F}_{n_j \times n_i}^t = \mathbf{A}_j^t \mathbf{A}_i \quad (\text{項目 } J \times I \text{ のクロス表})$$

レストランデータでは以下
 $N = 1,284$

$$n^* = n_i + n_j$$

\Updownarrow

$$n^* = n_i + n_j = 10 + 3 = 13$$

2項目の多重クロス表・バート表(B表)

対称行列	質問I レストラン	質問J 評価基準
質問I レストラン	(質問I) × (質問I) のクロス表 つまり質問Iの周辺度数が対角 要素に入った対角行列	(質問I) × (質問J) の クロス表 $\mathbf{F}_{n_i \times n_j} = \mathbf{A}_i^t \mathbf{A}_j$ ①
質問J 評価基準	(質問J) × (質問I) の クロス表 ② F表の転置行列 $\mathbf{F}_{n_j \times n_i}^t = \mathbf{A}_j^t \mathbf{A}_i$	(質問J) × (質問J) のクロス表 つまり質問Jの周辺度数が対 角要素に入った対角行列

固有値の記法の用意

- インジケータ行列(A表)とクロス表(F表)のCAから得られる固有値の関係を調べる.

⑤クロス表(F)の固有値

$$\lambda_k^F \begin{pmatrix} k = 1, 2, \dots, K \\ K = \min\{n_i, n_j\} - 1 \end{pmatrix}$$

⑥インジケータ行列(A)の固有値

$$\lambda_k^A \begin{pmatrix} k = 1, 2, \dots, K^* \\ K^* = n^* - M \\ n^* = n_i + n_j \end{pmatrix}$$

この両者の関係は？

ここで, M は項目数(いまは2)

2元データ表の相互比較(要約)

- 「第Ⅱ部」, 62ページあたりに要約の表を引用.
- とくに, “固有値”と“寄与率”の関係に注意する.

タイプ	データ表の形	データ表の次元数 (寸法)	固有値の関係
タイプ 1	<p>2元クロス表の場合</p> <p>2 項目 $I \times J$ または $J \times I$ のクロス表 $\mathbf{F}_{n_i \times n_j} = \mathbf{A}_i^t \mathbf{A}_j$ または $\mathbf{F}_{n_j \times n_i}^t = \mathbf{A}_j^t \mathbf{A}_i$</p> <p>(レストランデータの場合)</p>	$n_i \times n_j$ (*) 前に用いたクロス表の寸法を表す記号によると, $m = n_i, n = n_j$ と対応 (*) 固有値の個数は, $K = \min\{n_i, n_j\} - 1$	λ_k^F

$$n^* = n_i + n_j$$



$$n^* = n_i + n_j = 10 + 3 = 13$$

固有値の間の関係 (F表とA表の場合)

タイプ 2	2 項目 I と J のアイテム・カテゴリー型データ表 $\mathbf{A}_* = \begin{bmatrix} \mathbf{A}_i & \mathbf{A}_j \\ N \times n_i & N \times n_j \end{bmatrix}$ インジケータ行列 ここで $(n^* = n_i + n_j)$	$N \times n^* \quad (n^* = n_i + n_j)$	$\lambda_k^A = \frac{1 \pm \sqrt{\lambda_k^F}}{2}$
	タイプ 1 と タイプ 2 の固有値の関係: ここで, $n_i \geq n_j$ とする. <div style="float: right; color: blue;">(レストランデータの場合)</div> <div style="clear: both;"></div> <div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div> i) 値の大きい方から $n_j - 1$ 個 $\Rightarrow \lambda_k^A = \frac{1 + \sqrt{\lambda_k^F}}{2}$ ii) 値の小さい方から $n_j - 1$ 個 $\Rightarrow \lambda_k^A = \frac{1 - \sqrt{\lambda_k^F}}{2}$ iii) 間に含まれる $n_i - n_j$ 個 $\Rightarrow 1/2 = 0.5$ となる. (*) $n_i = n_j$ のときにはこれは現れない. </div> <div style="background-color: #e0ffe0; padding: 10px; border: 1px solid #c0ffe0;"> $\begin{cases} n_i = 10 > n_j = 3 \\ n_i - 1 \Leftrightarrow n_i - 1 = 9 \\ n_j - 1 \Leftrightarrow n_j - 1 = 2 \\ n_i - n_j \Leftrightarrow n_i - n_j = 7 \end{cases}$ </div> </div>		
タイプ 3 (2 項目 の場合)	2 項目の多重クロス表 (パート表) $\mathbf{B}_* = \mathbf{A}_*^t \mathbf{A}_* \quad (n^* = n_i + n_j)$ パート表	$n^* \times n^* \quad (n^* = n_i + n_j)$	$\lambda_k^B = (\lambda_k^A)^2 = \left(\frac{1 \pm \sqrt{\lambda_k^F}}{2} \right)^2$

クロス表(F表)とバート表(B表)のCA結果

(クロス表から出発の場合)

対応分析 特異値・固有値												
結果												
次元	特異値	固有値	割合(%)	.2	.4	.6	.8	累積(%)	.2	.4	.6	.8
1	0.44459	0.19766	76.7					76.7				
2	0.24498	0.06001	23.3					100				
固有値の合計 =0.257678804157674												

λ_k^F の確認

(バート表から出発の場合)

対応分析 特異値・固有値												
結果												
次元	特異値	固有値	割合(%)	.2	.4	.6	.8	累積(%)	.2	.4	.6	.8
1	0.72228	0.52169	18.1					18.1				
2	0.62223	0.38717	13.5					31.6				
3	0.5	0.25	8.7					40.3				
4	0.5	0.25	8.7					49.0				
5	0.5	0.25	8.7					57.7				
6	0.5	0.25	8.7					66.4				
7	0.5	0.25	8.7					75.1				
8	0.5	0.25	8.7					83.7				
9	0.49832	0.24832	8.6					92.4				
10	0.37707	0.14218	4.9					97.3				
11	0.27766	0.0771	2.7					100				
12	0.00039	1.55e-7	0.0					100				
固有値の合計 =2.8764664007726												

λ_k^B の確認

レストランの例でチェックする

$$\left[K^* = n^* - M \Leftrightarrow K^* = 13 - 2 = 11(\text{個}) \text{の固有値} \right]$$

(最後の1個は「0」, 計算誤差の範囲)

固有値の関係(F表とB表)

λ_k^B の表

成分	特異値	固有値	寄与率(%)	累積寄与率(%)
1	0.72228	0.52169	18.1	18.1
2	0.62223	0.38717	13.5	31.6
3	0.5	0.25	8.7	40.3
4	0.5	0.25	8.7	49
5	0.5	0.25	8.7	57.7
6	0.5	0.25	8.7	66.4
7	0.5	0.25	8.7	75.1
8	0.5	0.25	8.7	83.7
9	0.49832 (0.5)	0.24832 (0.25)	8.6	92.4
10	0.37707	0.14218	4.9	97.3
11	0.27766	0.0771	2.7	100

寄与率をどう読む？

λ_k^F の表

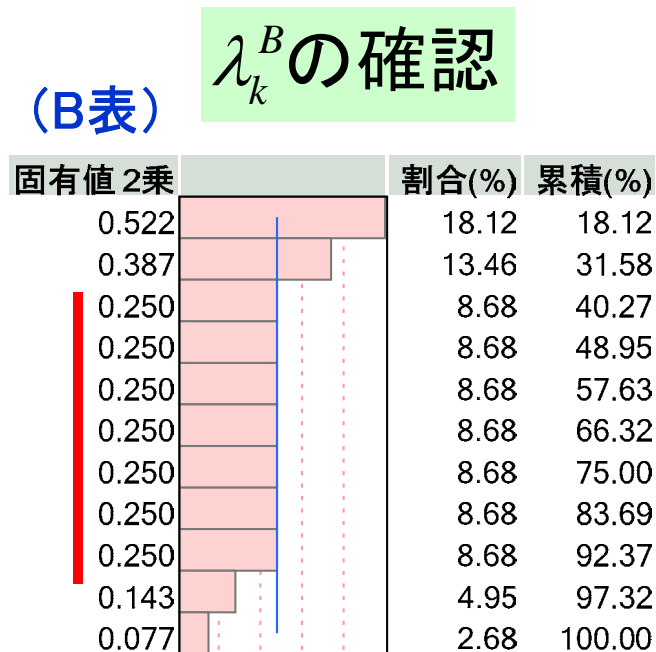
成分	特異値	固有値	寄与率(%)	累積寄与率(%)
1	0.44459	0.19766	76.7	76.7
2	0.24498	0.06001	23.3	100

$$\lambda_k^B = (\lambda_k^A)^2 = \left(\frac{1 \pm \sqrt{\lambda_k^F}}{2} \right)^2 \text{ の表}$$

成分	特異値	固有値	$\left(\frac{1 \pm \sqrt{\lambda_k^F}}{2} \right)^2$	
1	0.44459	0.19766	0.5217	0.0771
2	0.24498	0.06001	0.3875	0.1425

$\left(\begin{array}{l} \text{上下, } n_j - 1 = 3 - 1 = 2(\text{個}) \\ \text{の固有値; } n_i - n_j = 7 \text{個は} 0.5 \text{となる} \end{array} \right)$

インジケータ行列 (A表) とバート表 (B表)



レストランの例で確認する

- インジケータ行列の固有値の2乗が, バート表の固有値.
- 前ページの結果と合わせると, “2項目” の場合は以下の関係にある.

$$\lambda_k^B = (\lambda_k^A)^2 \quad \lambda_k^B = (\lambda_k^A)^2 = \left(\frac{1 \pm \sqrt{\lambda_k^F}}{2} \right)^2$$

複数の項目から得た多重クロス表

- 2項目を3項目以上, 一般にM項目に拡張しても事情はほぼ同じである.
- M 項目としたときのインジケータ行列(A表), バート表(B表)の関係は, 数理的に調べられている.
- A表のCAと, B表のCAとの式の導出と相互の関連は, テキスト, 第Ⅱ部に示した.
- 式の誘導はそれを確認いただくことにし, ここは結果の説明, 多重対応分析の仕組みを眺める.
- とくに注意すべきこととして, 両者の“固有値”と“寄与率”の関係がある. これを要約する.

(つづき)

- MCAについては、さまざまな研究がある。ここでは、そのごく一部を抜粋要約する。
- とくに、MCAに特有の扱いにくさがある。
- 項目の選択肢に付与の成分スコアと解釈、布置図の読み方。寄与度(絶対寄与度、相対寄与度・平方相関)の解釈など。
- ここでも、“はずれ値”や“度数の少ないセル”の影響を受けやすい。これへの手当の方法も考えられている。
- たとえば、該当行・列の一時除去、その再配置(追加処理を行う)など。

出発行列と慣性, 寄与率の関係

- おもな特性, それぞれの関係は数理的に調べられている.
- インジケータ行列 (A) の (列の) 大きさは, 項目数が増える
と急速に増える. (そして情報量が低減する)
- 固有値 (とその寄与率) は, 値が小さくなり, あたかも寄与
が低いように見える.
- これはデータ表の構造的な制約から生じるもの.
- 固有値と寄与率をそのまま読み取れない (解釈に難儀).
- これを回避・改善のために, たとえば“調整済み慣性”や
“調整済み寄与率” の提案がある. 後述.

調整済み総変動 (慣性) (adjusted inertia), 調整済み寄与率
(adjusted contribution)

ここで述べること

- インジケータ行列 (A) から出発した場合とバート表 (B) から出発した場合の, 固有値と総変動 (全慣性) の関係 を調べる.
 - ① インジケータ行列 (A) から出発したとき
 - ② バート表 (B) から出発したとき
 - ③ “調整済みの総変動 (全慣性)” と “調整済み寄与率”
- いろいろな研究があるが, ここでは, これらに限定して要約する.
- 簡単な数値例で確認する.

元の多変量構造データ(コーディング:C表)

表 32 一般的な調査データ: **C** 表(コーディング行列:大きさが $N \times n^*$)

回答者 \ 項目	Q_1	Q_2	...	Q_M
1	2	3	...	1
2	3	2	...	2
3	3	1	⋮	3
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
i	3	3	...	1
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
N	1	2	...	3
最大カテゴリ数	n_1	n_2	...	n_M

$\Rightarrow n^* = \sum_{j=1}^M n_j$

M 個の項目: $Q_1, Q_2, \dots, Q_j, \dots, Q_M$

各項目の選択肢数: $n_1, n_2, \dots, n_j, \dots, n_M$

インジケータ行列 (A表) とその行列記法

表 33 インジケータ行列: **A** 表

項目 回答者	Q_1	Q_2	...	Q_M	行和
1	0 1 0 0 ... 0	0 0 1 0 ... 0	...	1 0 0 0 ... 0	M
2	⋮	⋮	...	⋮	M
3	⋮	⋮	⋮	⋮	M
⋮	⋮	⋮	⋮	⋮	M
⋮	⋮	⋮	⋮	⋮	M
i	0 0 1 0 ... 0	0 0 1 0 ... 0	...	1 0 0 0 ... 0	M
⋮	⋮	⋮	⋮	⋮	M
⋮	⋮	⋮	⋮	⋮	M
N	1 0 0 0 ... 0	0 1 0 0 ... 0	...	0 0 1 0 ... 0	M
最大カテゴリー数	n_1	n_2	...	n_M	NM (総和)
総カテゴリー数	$n = n_1 + n_2 + \cdots + n_M = \sum n_j$				

(行和はすべて項目数: M)



$$\mathbf{A}_{N \times n^*} = \left[\begin{array}{c|c|c|c|c} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_i & \cdots & \mathbf{A}_j & \cdots & \mathbf{A}_M \\ \hline N \times n_1 & N \times n_2 & & N \times n_i & & N \times n_j & & N \times n_M \end{array} \right] \left(\text{ここで, } n^* = \sum_{j=1}^M n_j \right)$$

再確認&復習: 2元データ表の関係(3項目)

選択枝コードのデータ表(C)

サンプル	項目A1	項目A2	項目A3
1	1	1	1
2	1	1	2
3	1	2	1
4	2	1	2
5	2	1	2
6	2	1	3
7	2	2	1
8	2	2	2
9	3	1	2
10	3	1	3
11	3	1	3
12	3	2	1
13	3	2	1
14	3	2	2
15	3	2	2
16	3	2	3
17	4	1	3
18	4	2	2
19	4	2	2
20	4	2	3

C
20×3

$$\left(\begin{array}{l} M = 3 \\ n^* = 4 + 2 + 3 = 9 \end{array} \right)$$



インジケータ行列(A)

項目 選択枝	項目A1				項目A2		項目A3		
	1	2	3	4	1	2	1	2	3
1	1	0	0	0	1	0	1	0	0
2	1	0	0	0	1	0	0	1	0
3	1	0	0	0	0	1	1	0	0
4	0	1	0	0	1	0	0	1	0
5	0	1	0	0	1	0	0	1	0
6	0	1	0	0	1	0	0	0	1
7	0	1	0	0	0	1	1	0	0
8	0	1	0	0	0	1	0	1	0
9	0	0	1	0	1	0	0	1	0
10	0	0	1	0	1	0	0	0	1
11	0	0	1	0	1	0	0	0	1
12	0	0	1	0	0	1	1	0	0
13	0	0	1	0	0	1	1	0	0
14	0	0	1	0	0	1	0	1	0
15	0	0	1	0	0	1	0	1	0
16	0	0	1	0	0	1	0	0	1
17	0	0	0	1	1	0	0	0	1
18	0	0	0	1	0	1	0	1	0
19	0	0	0	1	0	1	0	1	0
20	0	0	0	1	0	1	0	0	1

(行和がすべて「3」になる)

3項目(M=3)からなるデータ表

A
20×9

多重クロス表（バート表）とインジケータ行列の関係

		項目A1				項目A2		項目A3		
		1	2	3	4	1	2	1	2	3
項目A1	1	3	0	0	0	2	1	2	1	0
	2	0	5	0	0	3	2	1	3	1
	3	0	0	8	0	3	5	2	3	3
	4	0	0	0	4	1	3	0	2	2
項目A2	1	2	3	3	1	9	0	1	4	4
	2	1	2	5	3	0	11	4	5	2
項目A3	1	2	1	2	0	1	4	5	0	0
	2	1	3	3	2	4	5	0	9	0
	3	0	1	3	2	4	2	0	0	6

$$\mathbf{A}^t \times \mathbf{A} = \mathbf{B}$$

9×20 20×9 9×9 バート表(B)

(項目) × (項目) の関係
[大きさが $n^* \times n^*$]

項目 選択肢	項目A1				項目A2		項目A3		
	1	2	3	4	1	2	1	2	3
1	1	0	0	0	1	0	1	0	0
2	1	0	0	0	1	0	0	1	0
3	1	0	0	0	0	1	1	0	0
4	0	1	0	0	1	0	0	1	0
5	0	1	0	0	1	0	0	1	0
6	0	1	0	0	1	0	0	0	1
7	0	1	0	0	0	1	1	0	0
8	0	1	0	0	0	1	0	1	0
9	0	0	1	0	1	0	0	1	0
10	0	0	1	0	1	0	0	0	1
11	0	0	1	0	1	0	0	0	1
12	0	0	1	0	0	1	1	0	0
13	0	0	1	0	0	1	1	0	0
14	0	0	1	0	0	1	0	1	0
15	0	0	1	0	0	1	0	1	0
16	0	0	1	0	0	1	0	0	1
17	0	0	0	1	1	0	0	0	1
18	0	0	0	1	0	1	0	1	0
19	0	0	0	1	0	1	0	1	0
20	0	0	0	1	0	1	0	0	1

$$\mathbf{A}$$

20×9 インジケータ行列表(A)

(回答者) × (項目) の関係
[大きさが $N \times n^*$]

この関係を用いるとバート表の対応分析で得た成分スコアの式を用いて、回答者(行)の成分スコアの算出が可能(後ろの分析例). 追加処理を使う.

バート表の構成要素(3項目のとき)

		項目A1				項目A2		項目A3		
		1	2	3	4	1	2	1	2	3
項目A1	1	3	0	0	0	2	1	2	1	0
	2	0	5	0	0	3	2	1	3	1
	3	0	0	8	0	3	5	2	3	3
	4	0	0	0	4	1	3	0	2	2
項目A2	1	2	3	3	1	9	0	1	4	4
	2	1	2	5	3	0	11	4	5	2
項目A3	1	2	1	2	0	1	4	5	0	0
	2	1	3	3	2	4	5	0	9	0
	3	0	1	3	2	4	2	0	0	6

		項目A1				項目A2		項目A3		
		1	2	3	4	1	2	1	2	3
項目A1	1	A1 × A1のクロス表 A1の周辺分布が対角				A1 × A2の クロス表		A1 × A3のクロス表		
	2									
	3									
	4									
項目A2	1	A2 × A1のクロス表				A2 × A2の クロス表		A2 × A3のクロス表		
	2									
項目A3	1	A3 × A1のクロス表				A3 × A2のク ロス表		A3 × A3のクロス表		
	2									
	3									

インジケータ行列の場合

- インジケータ行列(A)の対応分析で得られる総変動(全慣性)つまり固有値の総和.

①インジケータ行列 $\mathbf{A}_{N \times n^*}$ の固有値

$$\lambda_k^A \left(k = 1, 2, \dots, K^* \right) \left(\begin{array}{l} \text{ここで, } M \text{ は項目数, } n_j \text{ は項目 } j \text{ の選択肢数} \\ K^* = n^* - M, n^* = \sum_{j=1}^M n_j \end{array} \right)$$

②インジケータ行列の全慣性

$$inertia \left(\mathbf{A}_{N \times n^*} \right) = \frac{n^*}{M} - 1 = \frac{n^* - M}{M}$$

注:ここでいう固有値もすべて成分スコアの分散のこと

固有値の総和(全慣性):
$$\sum_k^{K^*} \lambda_k^A = \frac{n^*}{M} - 1$$

固有値判定の1つの目安

- インジケータ行列に展開したときの延べの次元数(n^*)の(項目数 M に対する)平均から1を引いた数に相当.
- 固有値 λ_k^A の平均は以下となる.
- インジケータ行列の場合, これを“目安”に固有値 λ_k^A が $1/M$ より大きいか, 小さいかを判断する(閾値).

③固有値の平均

$$\bar{\lambda}^A = \frac{1}{n^* - M} \sum_k^{K^*} \lambda_k^A = \frac{1}{n^* - M} \times \left(\frac{n^*}{M} - 1 \right) = \underline{\underline{\frac{1}{M}}}$$

例: レストランの例の場合, $M=2$, よって $1/M=1/2=0.5$
これを判断の目安とする. あとで例をみる.

固有値と寄与率

- ここで寄与率は以下ようになる.
- この寄与率には“上限がある”ことに注意.
- インジケータ行列の場合, 寄与率は⑤の式の右辺の値を越えることはない.

④固有値の寄与率

$$v_k = \frac{\lambda_k^A}{\sum_{k=1}^{K^*} \lambda_k^A} = \frac{\lambda_k^A}{\frac{n^*}{M} - 1} = \lambda_k^A \times \frac{M}{n^* - M} \quad (k = 1, 2, \dots, K^*)$$

⑤この寄与率の性質

$$v_k = \frac{\lambda_k^A}{\frac{n^*}{M} - 1} = \frac{M}{n^* - M} \lambda_k^A \leq \underline{\frac{M}{n^* - M}}$$

簡単な数値例(1)

例1: レストラン・データ

- $M=2$, $n^*=12$, よって $2/(12-2)=0.2$ (20%) を越えない.
- 下にみるように, 最大固有値の寄与率が13.1(%)である.

次元	固有値		割合(%)	累積(%)
1	0.722		13.133	13.133
2	0.622		11.318	24.451
3	0.500		9.091	33.542
4	0.500		9.091	42.632
5	0.500		9.091	51.723
6	0.500		9.091	60.814
7	0.500		9.091	69.905
8	0.500		9.091	78.996
9	0.500		9.091	88.087
10	0.378		6.864	94.951
11	0.278		5.049	100.000

$$\frac{1}{M} = \frac{1}{2} = 0.5 \text{ を目安とする}$$

簡単な数値例(2)

例2: テキスト, 第Ⅱ部にいくつか例を挙げた. 66ページから.

- いま, ある7つの質問項目を考える($M=7$). 各質問にはそれぞれ5つの選択肢があるものとする. ここで各固有値の寄与率は, 0.25(25%)を越えない.

$$\left\{ \begin{array}{l} M = 7, n^* = 5 \times 7 = 35 \\ K^* = n^* - M \Rightarrow 35 - 7 = 28 \\ \sum_{k=1}^{K^*} \lambda_k^A = \frac{n^*}{M} - 1 \Rightarrow \sum_{k=1}^{K^*} \lambda_k^A = \frac{n^*}{M} - 1 = \frac{28}{7} = 4 \end{array} \right. \Rightarrow \left\{ \begin{array}{l} \frac{M}{n^* - M} \\ = \frac{M}{K^*} = \frac{1}{4} = 0.25(25\%) \end{array} \right.$$

- (質問項目の総数)にくらべて、全質問の延べの総選択肢数が多くなると次第に固有値や寄与率が小さくなる.
 - つまり、項目の選択肢数を増やすほど情報が曖昧になる.
 - 形式的には (見かけ上は)すべての項目が“2項選択”がもつとも情報が多いとなる(そうみえる).
 - 不自然感を回避するため“調整済み慣性, 寄与率”を使う.
- 例3: M 個の項目のすべてが2項選択のとき

すべての項目の選択肢が2個のとき

$$n_j = 2 \quad (j = 1, 2, \dots, M), n^* = \sum_{j=1}^M n_j = 2M \text{ から}$$

$$\frac{M}{n^* - M} = \frac{M}{2M - M} = 1$$

バート表の固有値と総変動(全慣性)

- バート表(B)の多重対応分析で得られる固有値の和(全慣性)には以下の関係がある.
- また, インジケータ行列(A)の固有値との関係も下のようになる(すでに確認). 式のみ示す.

①バート表の全慣性(固有値の総和)

$$inertia(\mathbf{B}) = \sum_k^{K^*} (\lambda_k^A)^2 = \sum_k^{K^*} \lambda_k^B$$

②バート表の固有値とインジケータ行列の固有値の関係

$$(\lambda_k^A)^2 = \lambda_k^B \quad \text{または} \quad \lambda_k^A = \sqrt{\lambda_k^B}$$

“調整済みの総変動(全慣性)”(1)

- 前述のように, インジケータ行列(A)あるいはバート表(B)の固有値と寄与率には, 若干の不都合がある.
- 寄与率の解釈や選択のときに判断に迷う. この回避策.
- これを改善する試みがある. ここではBenzécriの基準とGreenacreの基準の2種を挙げる.

①Benzécriの調整済み慣性

$$\left\{ \begin{array}{l} \lambda_k^{adj} = \left(\frac{M}{M-1} \right)^2 \times \left(\lambda_k^A - \frac{1}{M} \right)^2 \\ \lambda_k^{adj} = \left(\frac{M}{M-1} \right)^2 \times \left(\sqrt{\lambda_k^B} - \frac{1}{M} \right)^2 \end{array} \right. (k = 1, 2, \dots, K^*)$$

- インジケータ行列 (A) の固有値の“平均”を用いると以下.
- とくに, 項目数 $M=2$ のとき, 前に確認の結果に一致.
- 2元クロス表 (F) の多項目への拡張, 一般化になっている.

②2元クロス表のときとの関係

$$\lambda_k^{adj} = \left(\frac{M}{M-1} \right)^2 \times (\lambda_k^A - \bar{\lambda}^A)^2 \quad (k = 1, 2, \dots, K^*)$$

とくに, $M = 2$ のとき

$$\lambda_k^{adj} = 4 \left(\sqrt{\lambda_k^B} - \frac{1}{2} \right)^2 = 4 \left(\lambda_k^A - \frac{1}{2} \right)^2 \quad (\equiv \lambda_k^F)$$

λ_k^A について解くと以下となる.

$$\lambda_k^A = \frac{1}{2} \left(1 \pm \sqrt{\lambda_k^{adj}} \right) \left[\equiv \frac{1}{2} \left(1 \pm \sqrt{\lambda_k^F} \right) \right]$$

前にみた結果に同じ
スライド, 24~27ページあたり

“調整済みの総変動(全慣性)”(2)

- Greenacreは以下のような基準を提案した.
- これを, “非対角の慣性(2乗和)の平均”(average off-diagonal inertia)という. 適当な訳語がないのでこうした.

①非対角の慣性(2乗和)の平均

$$\frac{M}{M-1} \times \left(inertia(\mathbf{B}) - \frac{n^* - M}{M^2} \right)$$

②非対角の2乗和と対角の2乗和

$$\text{非対角の2乗和} \quad inertia(\mathbf{B}) - \frac{n^* - M}{M^2}$$

$$\text{対角の2乗和} \quad \frac{n^* - M}{M^2}$$

意識調査データを用いた分析例

- 実際の調査データを用いたMCAによる分析例をみる.
- 元のデータセットの質問項目数はかなり多い, 内容も多岐にわたるので, 再編集した例を用いる.
- 調査の背景, 構成概念などが分からなくても直感的に理解できそうな複数の質問と, 分析軸の組み立てに役立つ客観的情報として人口統計学的変数のいくつかを加えた.
- ここで, MCAによる分析にあたって留意すべき事項を要約する. ここは「★参考」としておこう. 演習時のヒントとしていただくとよいだろう.

★参考:MCAによる分析時の留意事項

- 一般に, 複数の項目の多重対応分析は複雑になる.
- まず, とりあげる質問項目の“**初動探査**”を行う.
- 個々の質問項目ごとの回答分布の傾向の観察, 無回答や欠測値の確認する.
- とくに, 2元データ表の周辺和(行和, 列和)の頻度の小さい選択肢には注意(また, 質量を観察). はずれ値となりやすい.
- 2つの質問項目間の探査, **クロス表分析**などの観察.
- “**独立性の検定**”の結果も参考にする.

(つづき)

- 必要に応じて、データの“簡易編集・加工”を行う。たとえば、無回答、欠測値の除去、回答分布を観察し偏った選択肢を併合するなど。
- MCAを利用するときに、いくつかの“分析シナリオ”(分析要領のメモ)を用意するとよいだろう。
- このとき、質問項目と人口統計学的変数や、可能であれば外部情報源情報の利用を想定する。これは解釈を客観化するための基本操作である。

(つづき)

- “**仮説発見的**”に、複数の質問項目の関係を探索する.
- “**1つの質問項目**”と人口統計学的変数などの客観情報 (できれば複数)を組み合わせて探索.
- ここで、初動探索でえた情報を勘案し選び方を決める.
- つぎに、はじめにみた“**複数の質問項目**”と“人口統計学的”変数ないしは類似の外部情報源を指定する.
- 回答の揺らぎがない(実態をみる)変数を加える.
- こうした操作を繰り返して、項目の組み合わせで見えてくる傾向を拾い出す. “**傾向探索**”, “**特徴抽出**”を行う.

(つづき)

- 布置図・同時布置図は，探査・発見的に観察するための“思考地図”(意識マップ)を描くことに相当する.
- しかし多次元情報であり，情報の一部をみていることにも注意.
- 同時に，成分スコアは“合成変数”でもあることも想起.
- なんども用いた「環境意識調査」や「自治体市民意識調査」の再加工データを用いて，簡単な分析を試みる.

(つづき)

- これらは演習用サンプル・データとして用意したので、各自で試みていただくとよいだろう.
- 断り:MCAが何を行うかの見通しをよくするために、内容はかなり簡略化されている. 現実の調査データは、より一層複雑である.
- たとえば, この資料のうしろに付けた**ピエール・ブルデューのマップ**を参照. かなり複雑である.

例1:「環境意識調査」データを用いた例

- なんども用いたこのデータの、3カ年, 6調査地域の全体に共通して用いた質問項目のいくつかを選ぶ.
- 選んだ質問項目は以下の4つ. 選択肢数に注意する.
 - Q1(2):あなたは、いま住んでいるまちが気に入っていますか. (4選択肢)
 - Q2(1):住んでいる地区は、都市としては、緑(みどり)が多いと感じますか. (5選択肢)
 - Q9:できることなら、今後とも現在のまちに住みたい、と思いますか. (4選択肢)
 - Q10(1):あなたの生活状態は、この10年間でよくなりましたか. (5選択肢)

(つづき)

- これと、人口統計学的変数として「性別(2選択肢)」「年齢区分(7区分)」「性年齢区分(14区分)」がある.
- つまり、質問文以外の項目として以下がある(要約). これらは、考察時の客観化情報として補完的に利用する.
 - 調査地点: 6地点(抽出時に確定するので無回答はなし)
 - 性別: 2選択肢(無回答は一時除去)
 - 年齢区分: 7選択肢(実年齢と回答者記入とから再編集)
- 報告で用いたバート表を示すが、ここで用意した再編集データは、(無回答処理などで)若干内容を変えてある.

分析のシナリオ

- 用いる項目の組合せをいくつか用意する(シナリオ).
- 本来, 調査設計時に検討すべきことである. 設定された質問文の検証でもある.
- 見方はさまざまである. よってシナリオもいろいろある.

分析1: $[M = 4, K^* = n^* - M = 20 - 4 = 16]$

- Q2(1): 住んでいる地区は, 都市としては, 緑(みどり)が多いと感じますか. (5選択肢) [緑が多いか]
- 「調査地域」(6地点), 人口統計学的変数として「性別」(2選択肢), 「年齢区分」(7選択肢)
- 「緑の多さ」(への意識)は, 地域とどう関連するか. 回答者属性は関係するのか, しないのか.

質問項目を増やすと, ...

分析2: $[M = 4, K^* = n^* - M = 14 - 4 = 10]$

- 質問項目(選択肢型)を増やして, 以下の3項目とする.
 - あなたは, いま住んでいるまちが気に入っていますか. (4選択肢)
 - 住んでいる地区は, 都市としては, 緑(みどり)が多いと感じますか. (5選択肢)
 - あなたの生活状態は, この10年間でよくなりましたか. (5選択肢)
- 対比」させる項目は「調査地域」(5地域)の1項目とする.
- 人口統計学的変数は除外する.
- 各質問項目間の関連を調べる同時に, 調査地域との関係を探査する.
- ここで, 「調査地域」をはずして, 人口統計学的変数と比べるとどうなるだろう. 演習問題としておこう,

いきなり項目数を増やすと, ...

分析3: $[M = 7, K^* = n^* - M = 33 - 7 = 26]$

- 質問項目(選択肢型)をすべて, 4項目とする.
 - Q1(2): あなたは, いま住んでいるまちが気に入っていますか. (4選択肢)
 - Q2(1): 住んでいる地区は, 都市としては, 緑(みどり)が多いと感じますか. (5選択肢)
 - Q9: できることなら, 今後とも現在のまちに住みたい, と思いますか. (4選択肢)
 - Q10(1): あなたの生活状態は, この10年間でよくなりましたか. (5選択肢)
- 対比」させる項目は以下の3項目とする.
 - 「調査地域」(6地点), 人口統計学的変数として「性別」(2選択肢), 「年齢区分」(7選択肢)

分析1:バート表の例(性別, 年齢区分, 地域, 「緑が...」)

表 IV.10 パート表

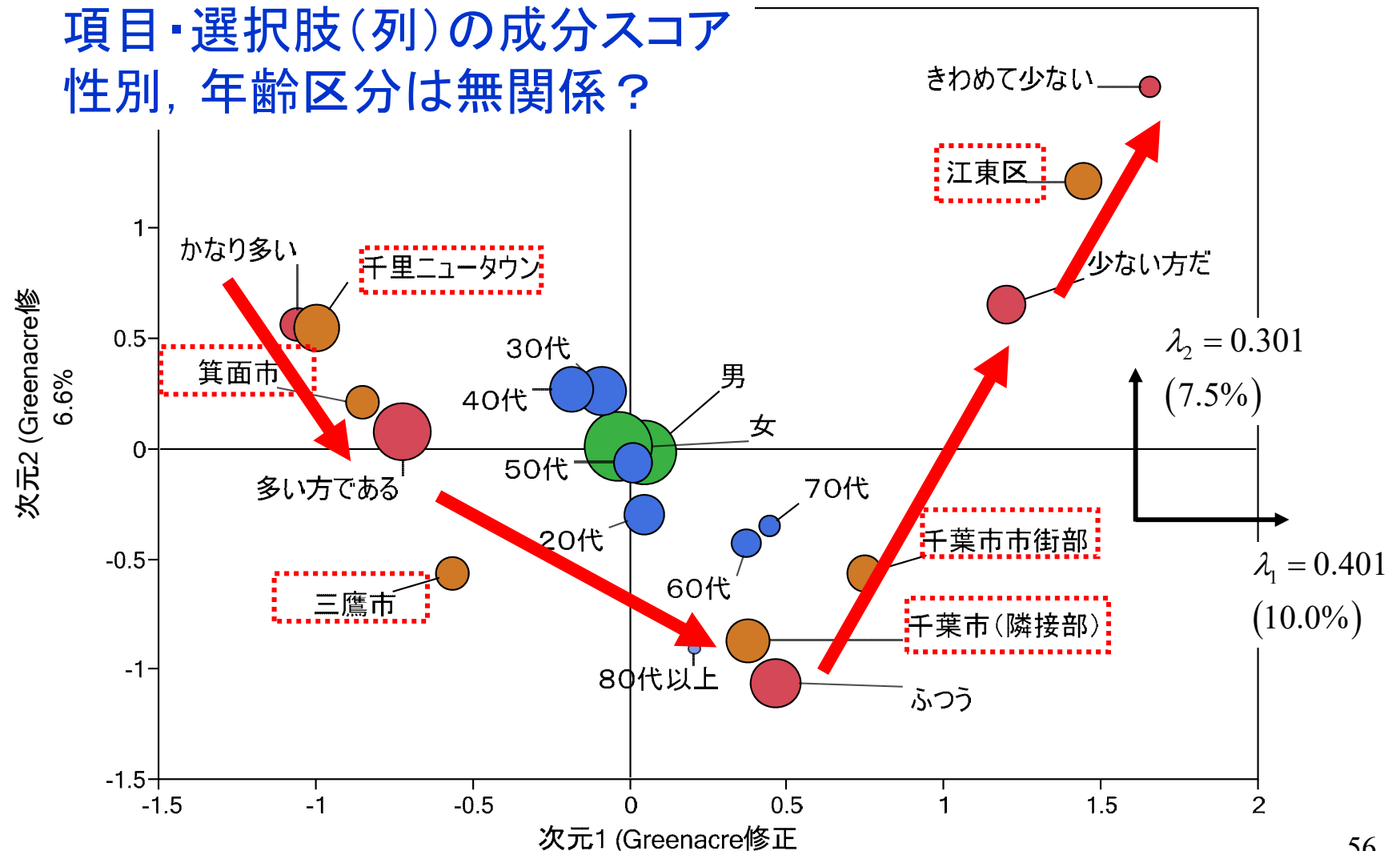
		性 別		年 齢 区 分							地 域						質 問 B						
		男	女								1千里 タウニ ン	千市 葉街 部	箕面 市	千葉 市 隣接 部	江東 区	三鷹 市	いかなり 多	ある方 で	多い方 で	ふつう	だ少ない 方	少ない 方	きわめて 少ない DK
				20代	30代	40代	50代	60代	70代	80代以上													
性別	男 子 女 子	2329 0	02647	505 516	616 723	524 504	359 406	189 224	102 117	34 57	552 650	355 413	275 360	512 562	345 373	290 289	308 360	849 1007	640 742	425 396	108 131	4 11	
年 齢 区 分	20代	505	516	1021	0	0	0	0	0	0	232	129	137	234	151	138	123	381	292	182	40	3	
	30代	616	723	0	1339	0	0	0	0	0	348	181	187	297	187	139	179	527	342	230	57	4	
	40代	524	504	0	0	1028	0	0	0	0	328	165	137	232	152	113	154	454	284	175	58	2	
	50代	359	406	0	0	0	765	0	0	0	173	127	91	165	111	98	116	256	234	109	37	3	
	60代	189	224	0	0	0	0	413	0	0	69	93	43	83	70	55	48	138	136	69	21	1	
	70代	102	117	0	0	0	0	0	119	0	25	45	30	50	42	27	35	67	69	35	12	1	
	80代以上	34	57	0	0	0	0	0	0	91	27	28	10	12	5	9	13	33	25	16	4	0	
地 域	千里ニュータウン	552	650	232	348	328	173	69	25	27	1202	0	0	0	0	0	281	709	163	42	4	6	
	千葉市市街部	355	413	129	181	165	127	93	45	28	0	768	0	0	0	0	26	179	306	187	65	5	
	箕面市	275	360	137	187	137	91	43	30	10	0	0	635	0	0	0	176	312	115	30	2	0	
	千葉市隣接部	512	562	234	297	233	165	83	50	12	0	0	0	1074	0	0	60	314	442	213	45	0	
	江東区	345	373	151	187	152	111	70	42	5	0	0	0	0	718	0	15	60	211	308	122	2	
	三鷹市	290	289	138	139	113	98	55	27	9	0	0	0	0	0	579	111	282	147	36	1	2	
質 問 B	かなり多い	308	360	123	179	154	116	48	35	13	281	26	176	60	15	111	668	0	0	0	0	0	
	多い方である	849	1007	381	527	454	256	138	67	33	709	179	312	314	60	282	0	1856	0	0	0	0	
	ふつう	640	742	292	342	284	234	136	69	25	163	306	115	442	211	147	0	0	1382	0	0	0	
	少ない方だ	425	396	182	230	175	109	69	35	16	42	187	30	213	308	36	0	0	0	0	0	0	
	きわめて少ない	108	131	40	57	58	37	21	12	4	4	65	2	45	122	1	ある記述から(大						
	DK	4	11	3	4	2	3	1	1	0	6	5	0	0	2	2	0	0	0	0	0	13	

「の程度で」「のサイズ

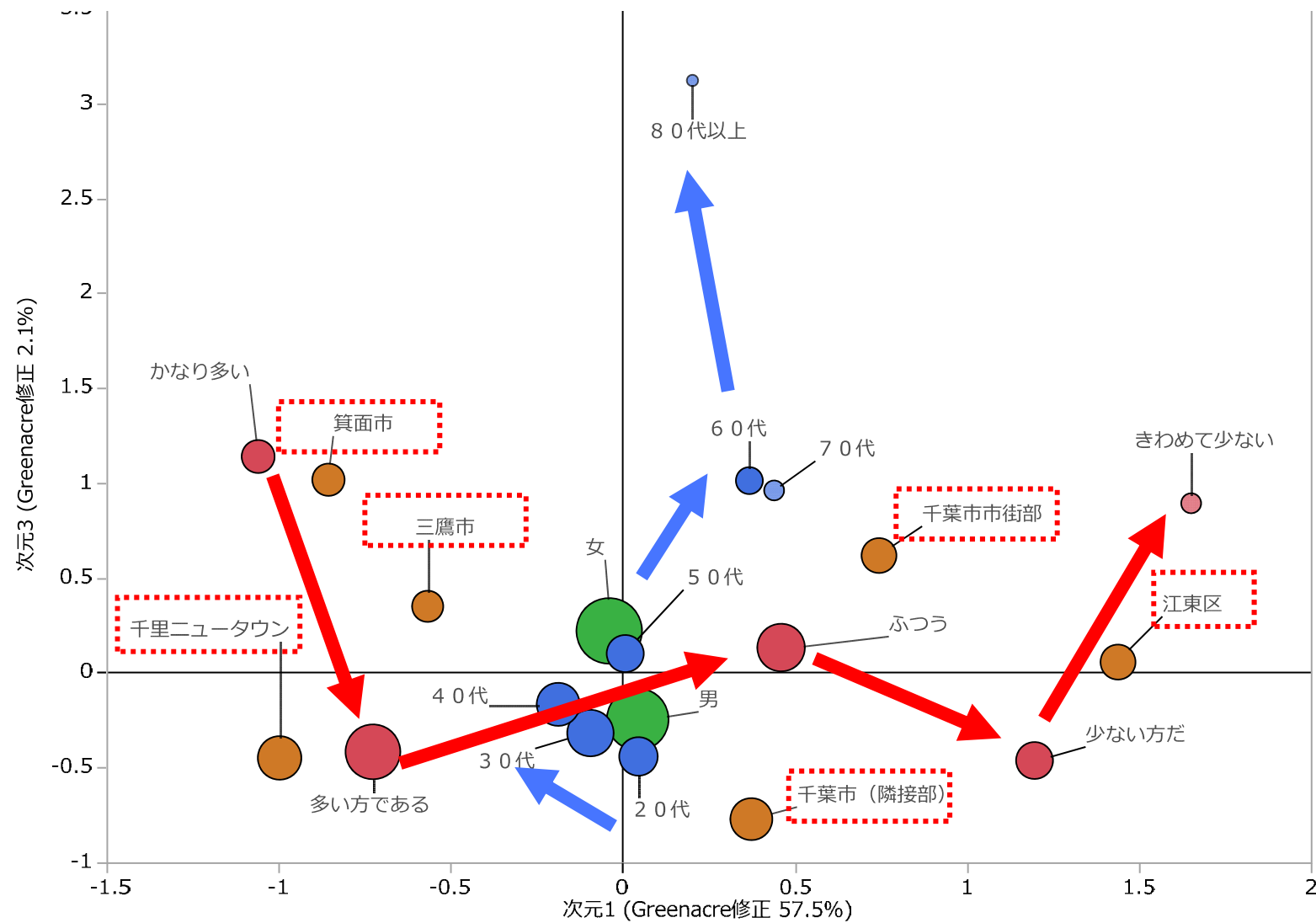
この下に「(回答者) × (0, 1)パターン」のインジケータ行列を想定

同じ分析を追試, まず(1, 2)成分で観察

項目・選択肢(列)の成分スコア
性別, 年齢区分は無関係?



(1, 3)成分で観察する

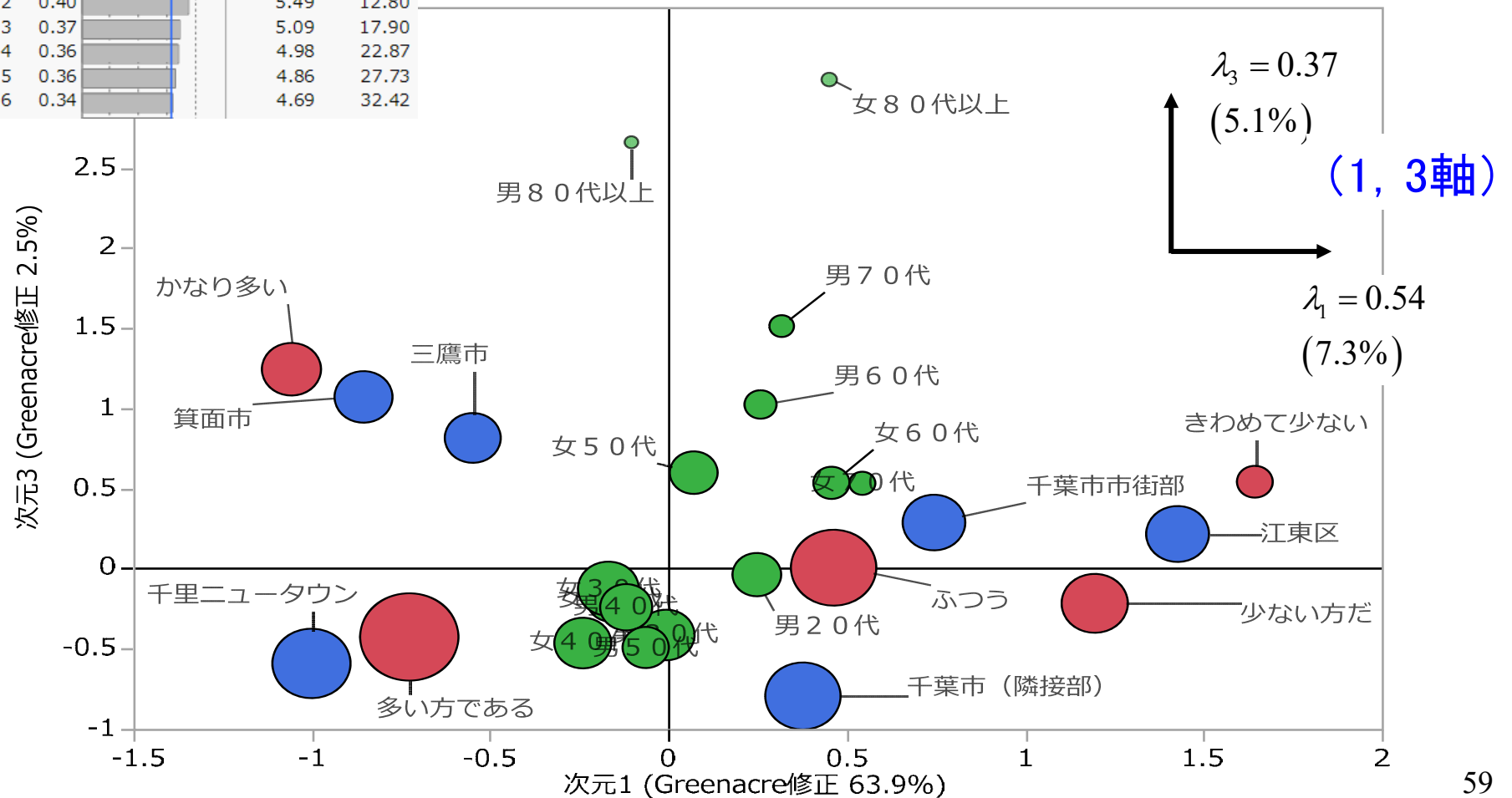


参考:

- ここで, 人口統計学的変数で, 「性別」「年齢区分」を括って「性年齢区分」(14選択肢)とするとどうなるか.
- Q2(1): 「緑が多いか」(5選択肢)
- 「調査地域」(6地点)
- 次ページに(1, 3)成分の同時布置図を表示した.
- 性年齢区分としたことは, 調査地域と「緑が多いか」の関係におおきくは響かないようだ(性年齢区分で分析してよさそう).
- いずれも, 性別, 年齢区分は, 別の成分に関与(第3成分)している.
- 確認

(1, 3)成分で比較, 観察する

次元	固有値	固有値	割合(%)	累積(%)
1	0.54		7.31	7.31
2	0.40		5.49	12.80
3	0.37		5.09	17.90
4	0.36		4.98	22.87
5	0.36		4.86	27.73
6	0.34		4.69	32.42



これを観察して, ...

- 寄与率の大きさに注意. みたところかなり小さいが, これをどう読むのか. 次ページから.
- この次元内では, 人口統計学的変数の, 性別, 年齢区分は中央に分布している(標本の偏りと高年齢層の無回答).
- 用いた質問項目(緑が...)と調査地域の関係が支配的である.
- ここで用いた多重クロス表(バート表)の寄与率をどう観察するか.
- 標準的な固有値, 寄与度と, 調整済み寄与率を比べて評価する.

インジケータ行列とバート表の固有値ほか

λ_k^A の確認

λ_k^B の確認

次元	固有値		割合(%)	累積(%)	固有値 2乗		割合(%)	累積(%)
1	0.401		10.03	10.03	0.161		15.28	15.28
2	0.301		7.53	17.55	0.091		8.61	23.88
3	0.279		6.97	24.53	0.078		7.39	31.28
4	0.274		6.84	31.37	0.075		7.12	38.40
5	0.261		6.53	37.90	0.068		6.48	44.87
6	0.256		6.40	44.30	0.066		6.23	51.10
7	0.253		6.32	50.62	0.064		6.06	57.16
8	0.250		6.26	56.88	0.063		5.95	63.11
9	0.249		6.23	63.11	0.062		5.90	69.01
10	0.246		6.14	69.24	0.060		5.73	74.74
11	0.238		5.95	75.19	0.057		5.37	$\frac{1}{M} = \frac{1}{4} = 0.25$
12	0.236		5.90	81.09	0.056		5.29	\Downarrow
13	0.228		5.71	86.80	0.052		4.95	$\lambda_k^B = (\lambda_k^A)^2$
14	0.226		5.64	92.44	0.051		4.84	$\lambda_k^B \geq \left(\frac{1}{4}\right)^2 = 0.0625$
15	0.201		5.03	97.46	0.040		3.84	
16	0.101		2.54	100.00	0.010		0.98	

ここを目安, 青い縦線に同じ

$$\frac{1}{M} = \frac{1}{4} = 0.25$$

$$\Downarrow$$

$$\lambda_k^B = (\lambda_k^A)^2$$

$$\lambda_k^B \geq \left(\frac{1}{4}\right)^2 = 0.0625$$

調整済み固有値 (Benzécri, Greenacre)

- これによると, はじめの2~3成分に注目すればよさそう.
- この2つの指標は, 若干, 情報量のはしよりすぎとなる傾向にある.

調整固有値(2乗)	Benzécri 割合(%)	Benzécri	Benzécri 累積(%)	Benzécri	Greenacre 割合(%)	Greenacre	Greenacre 累積(%)	Greenacre
0.041	84.53		84.53		57.53		57.53	
0.005	9.64		94.17		6.56		64.09	
0.001	3.11		97.28		2.12		66.20	
0.001	2.09		99.38		1.43		67.63	
0.000	0.46		99.84		0.31		67.94	
0.000	0.14		99.97		0.09		68.04	
0.000	0.03		100.00		0.02		68.06	
0.000	0.00		100.00		0.00		68.06	
.	
.	
.	
.	
.	
.	
.	
.	

回答者の成分スコアの観察

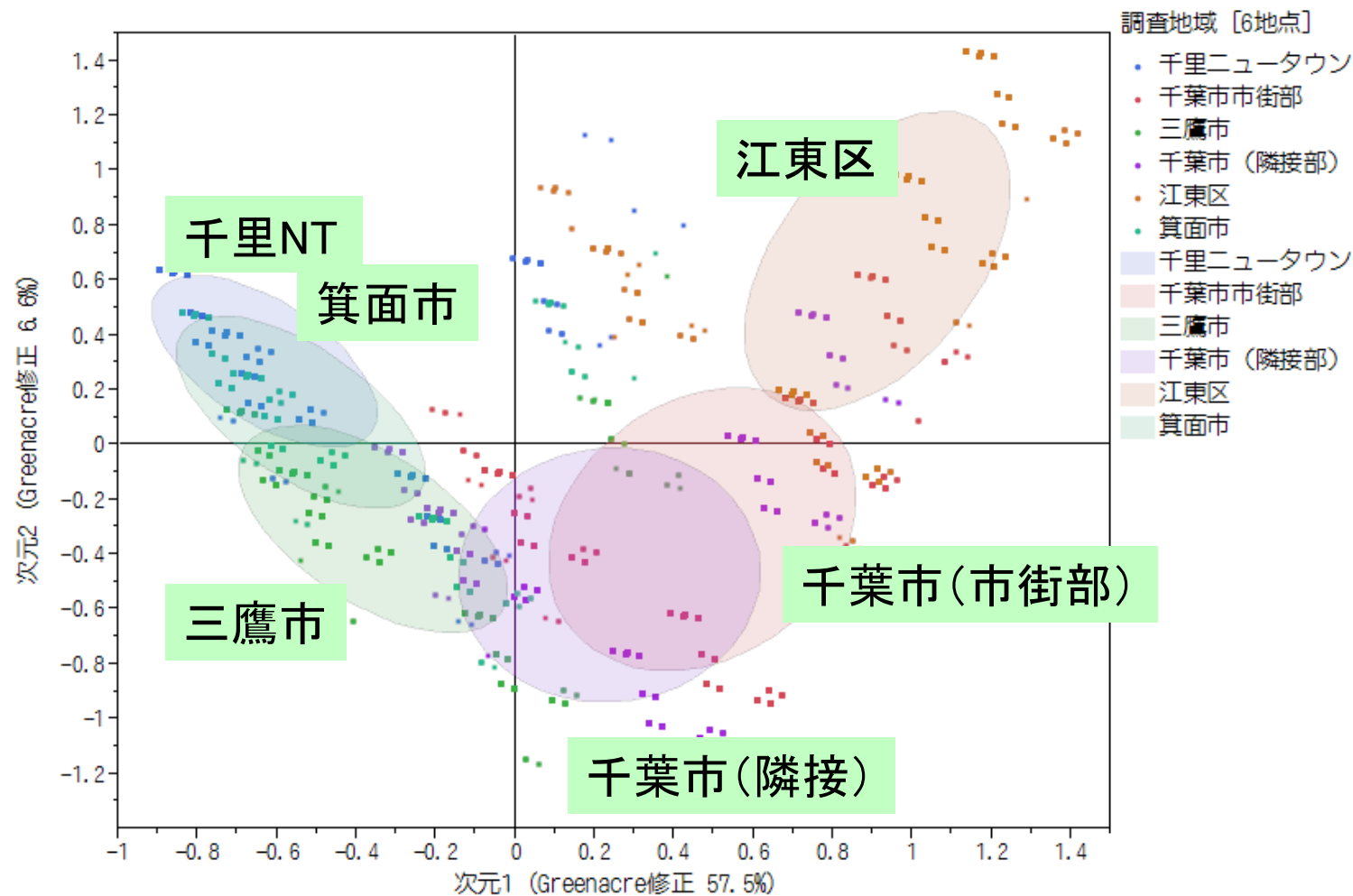
- 手順の詳しいことは略して, こんなことを行うことができる. と例でフォローする.
- バート表で得た成分式に追加処理を行うことで回答者側の成分スコアが得られる (前にイメージで確認した
- 「(回答者) \times (0, 1) パターン」のインジケータ行列を追加すること.
- 得られた成分スコアの布置図を描き, 用いた質問項目の選択肢別に“確率楕円”を描いて, 意見分布の様子を観察する.

回答者の成分スコアの観察

- 各確率楕円(アミ)は95%信頼限界(カイ二乗距離)に相当する(詳しい説明は略す).
- 確率楕円の重心と囲みが離れていれば, 選択肢間の差異がある. 重なれば差異が少ない, と読む.
- 前にみた選択肢側の同時布置図と併せて観察する.
- 各回答者の応答は, インジケータ行列上で(0, 1)と対応していることを念頭に, ごくおおまかに各項目の回答者パターンを観察する.

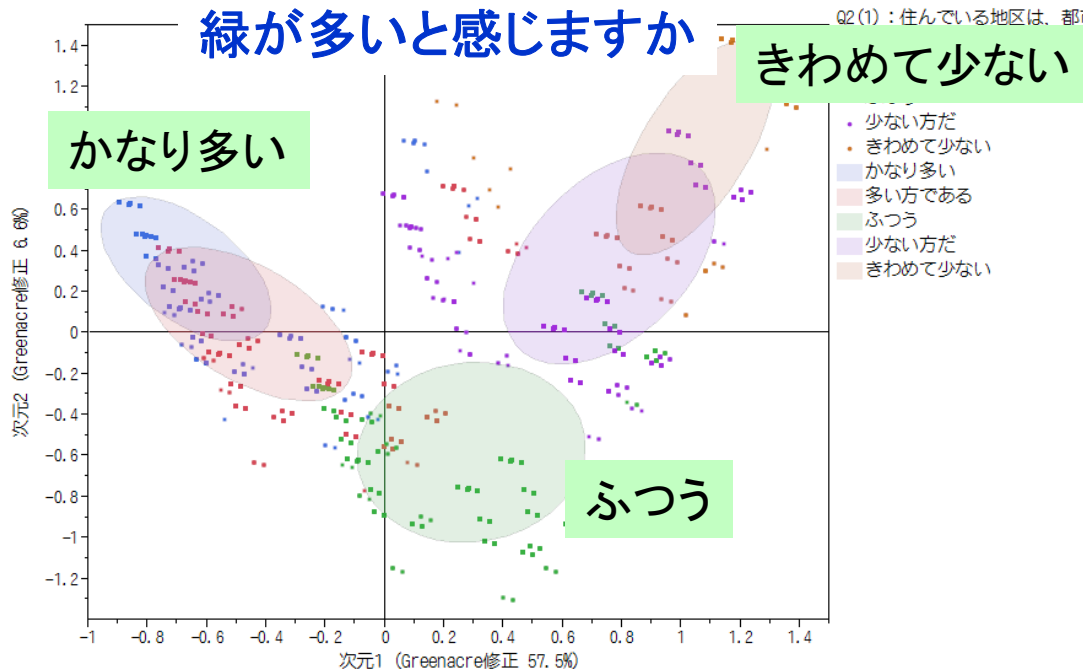
回答者(行)の成分スコアの分布(調査地域)

行座標の布置図

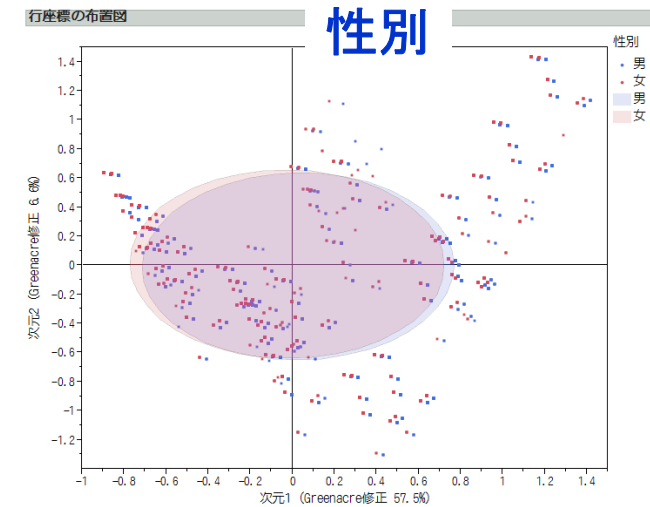


質問「緑が多いか...」と「性別」「年齢区分」

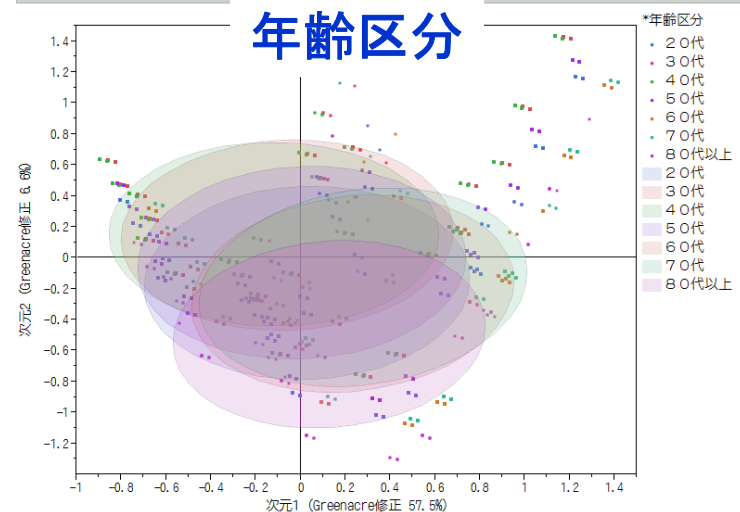
行座標の布置図



行座標の布置図

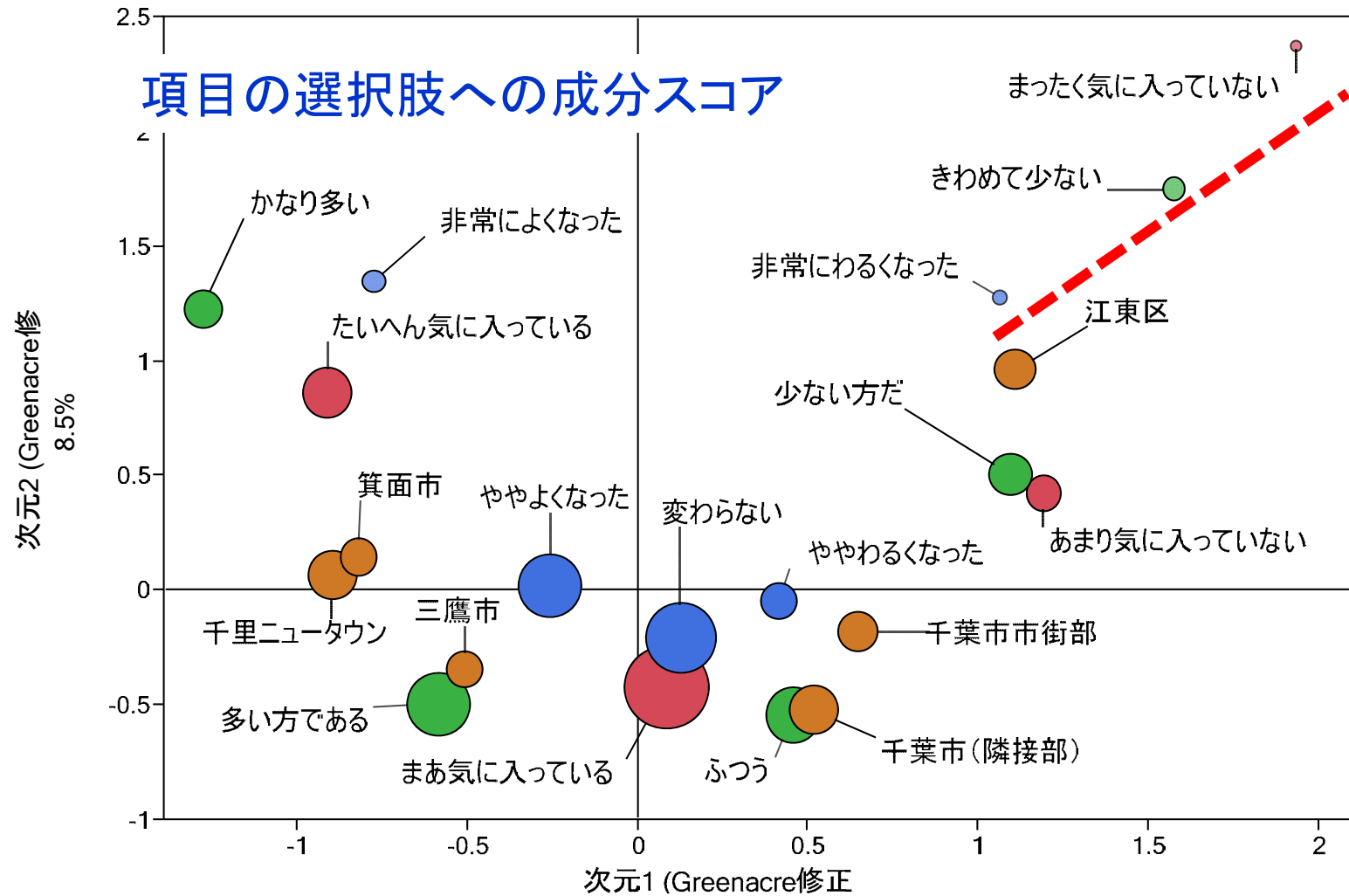


行座標の布置図



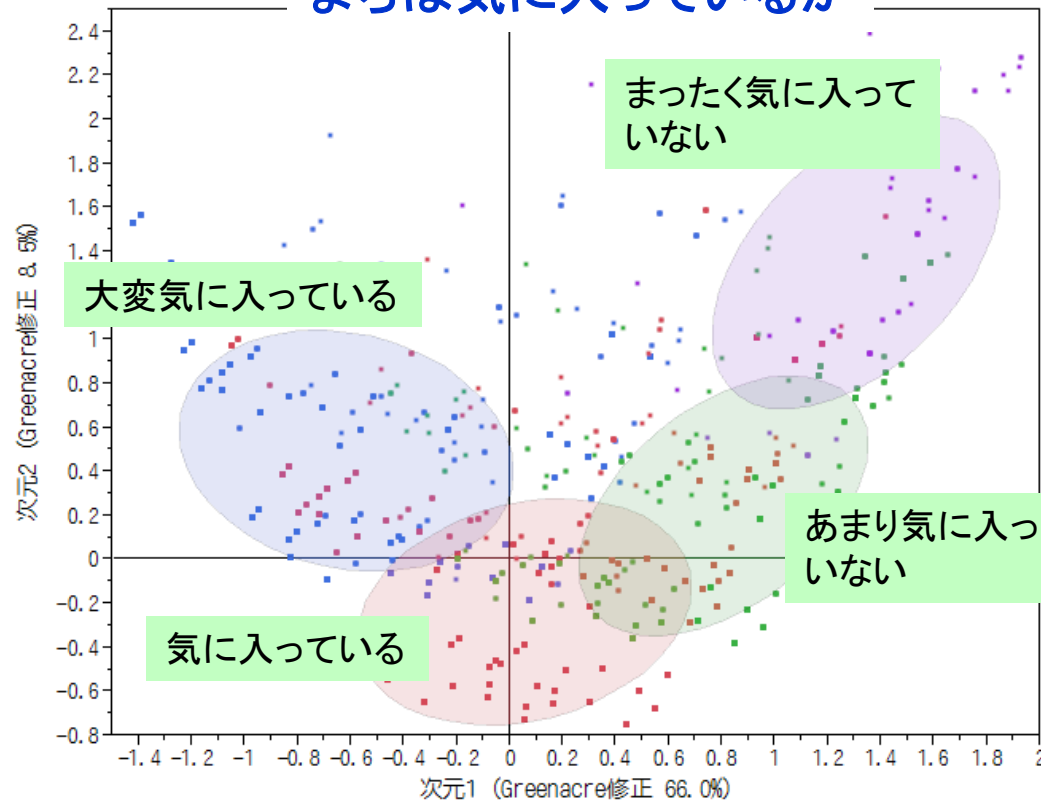
- 「緑が...」は「調査地域」と関連あり、常識的な結果.
- 「性別」「年齢区分」は、別の次元、調査地域とは関連はなさそう.

分析2: 項目数を増やす(3項目 + 調査地域)



回答者(行)の成分スコア

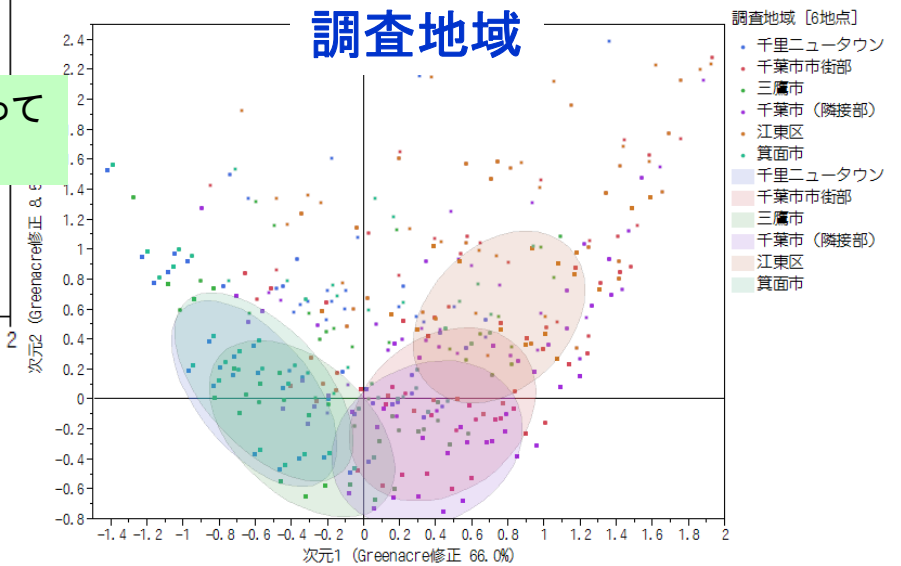
まちは気に入っているか

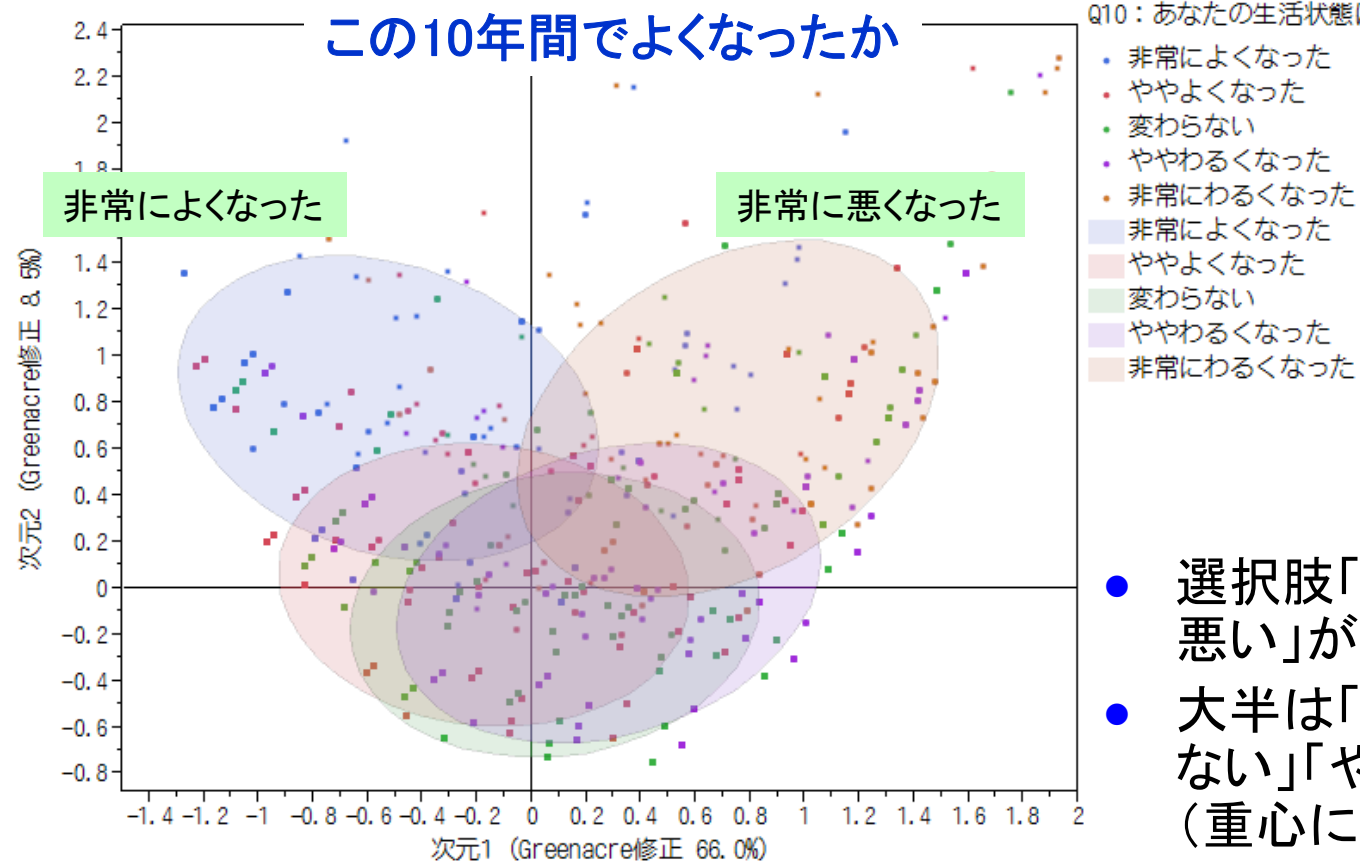


Q1 (2) : あなたは、いま住んでいるまちが気に入っていますか。

- ・ たいへん気に入っている
- ・ まあ気に入っている
- ・ あまり気に入っていない
- ・ まったく気に入っていない
- ・ たいへん気に入っている
- ・ まあ気に入っている
- ・ あまり気に入っていない
- ・ まったく気に入っていない

調査地域



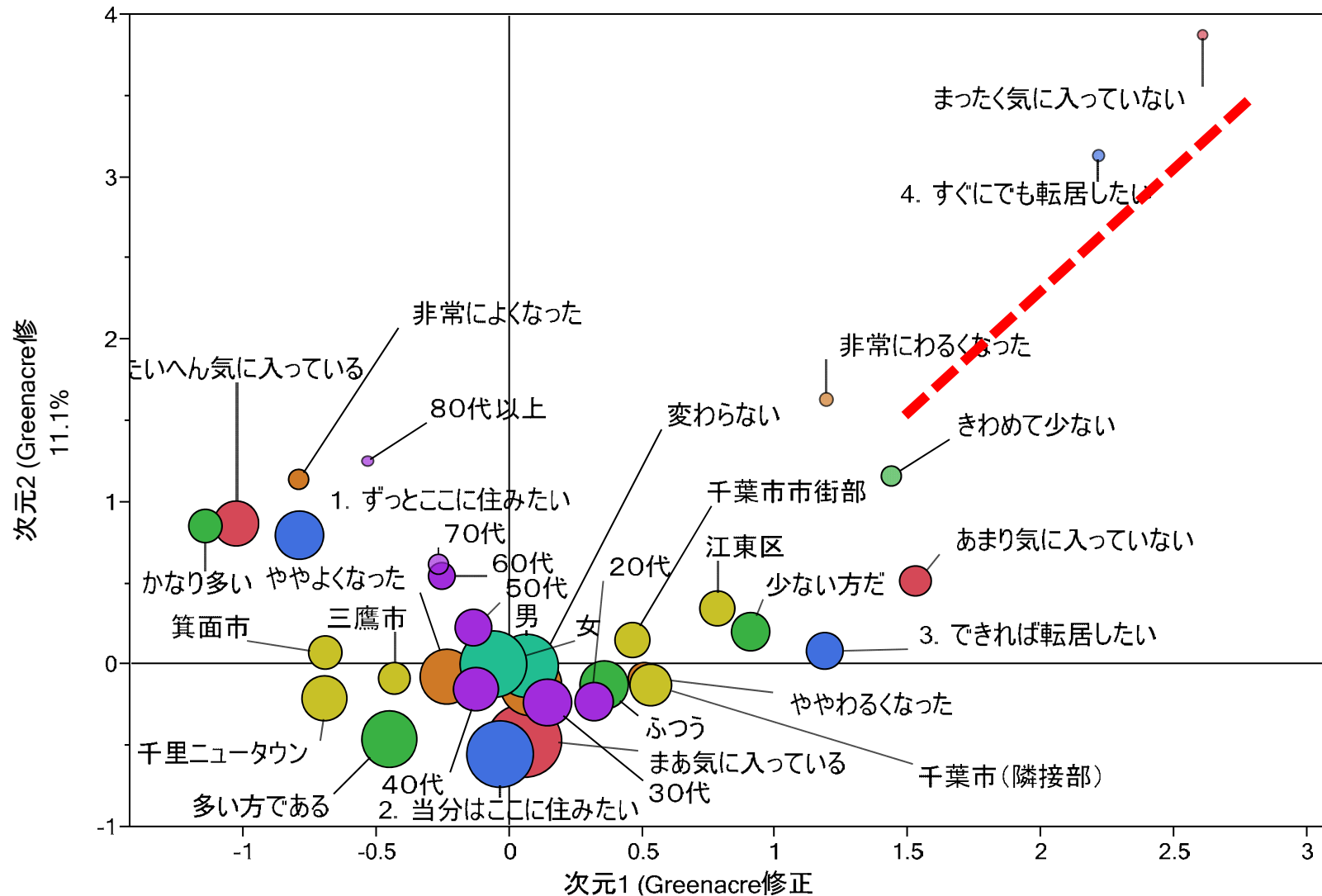


- 選択肢「非常よい」「非常に悪い」が異なる意見.
- 大半は「ややよい」「変わらない」「やや悪い」, 平均的 (重心に近い)

分析2の演習課題へのメモ

- 「調査地域」をはずして、人口統計学的変数と比べるとどうなるか.
- たとえば、「性別」「年齢区分」を使うとどうか.
- 「性年齢区分」を使うとどうなるか.
- ここで、「性別」と「年齢区分」の調査地域別のクロス表分析に注意のこと.
- これを行うと、これら“人口統計学的変数”が、3つの質問項目とさほど高い関連にあるようにはみえないことが観察される.
- つまり、この3つの質問については、調査地域間の差違が顕著で、人口統計学的変数の違いは少ないのではないかとみられること。(しかし、単純ではない)

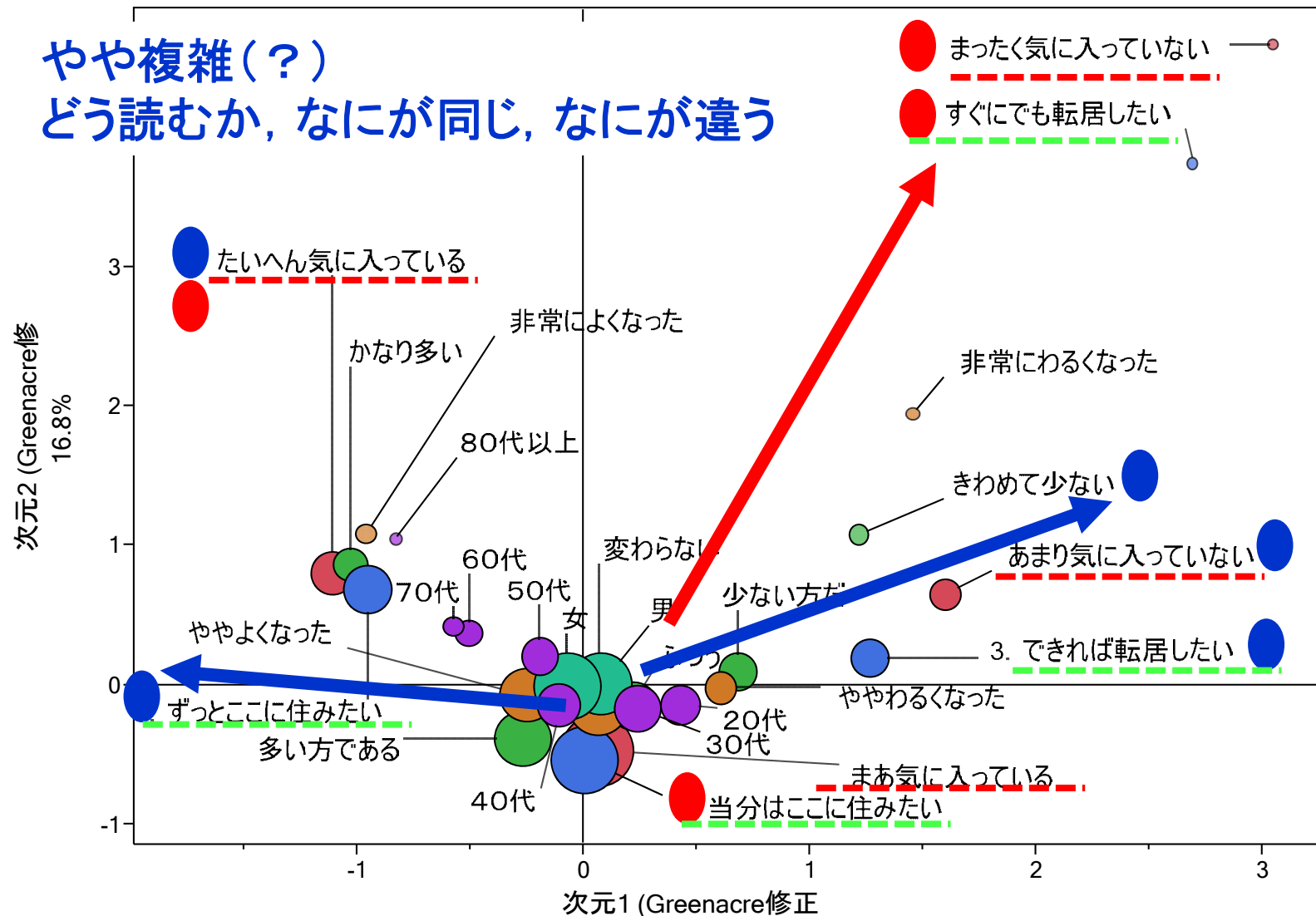
分析3:すべての項目を利用, どう読む?



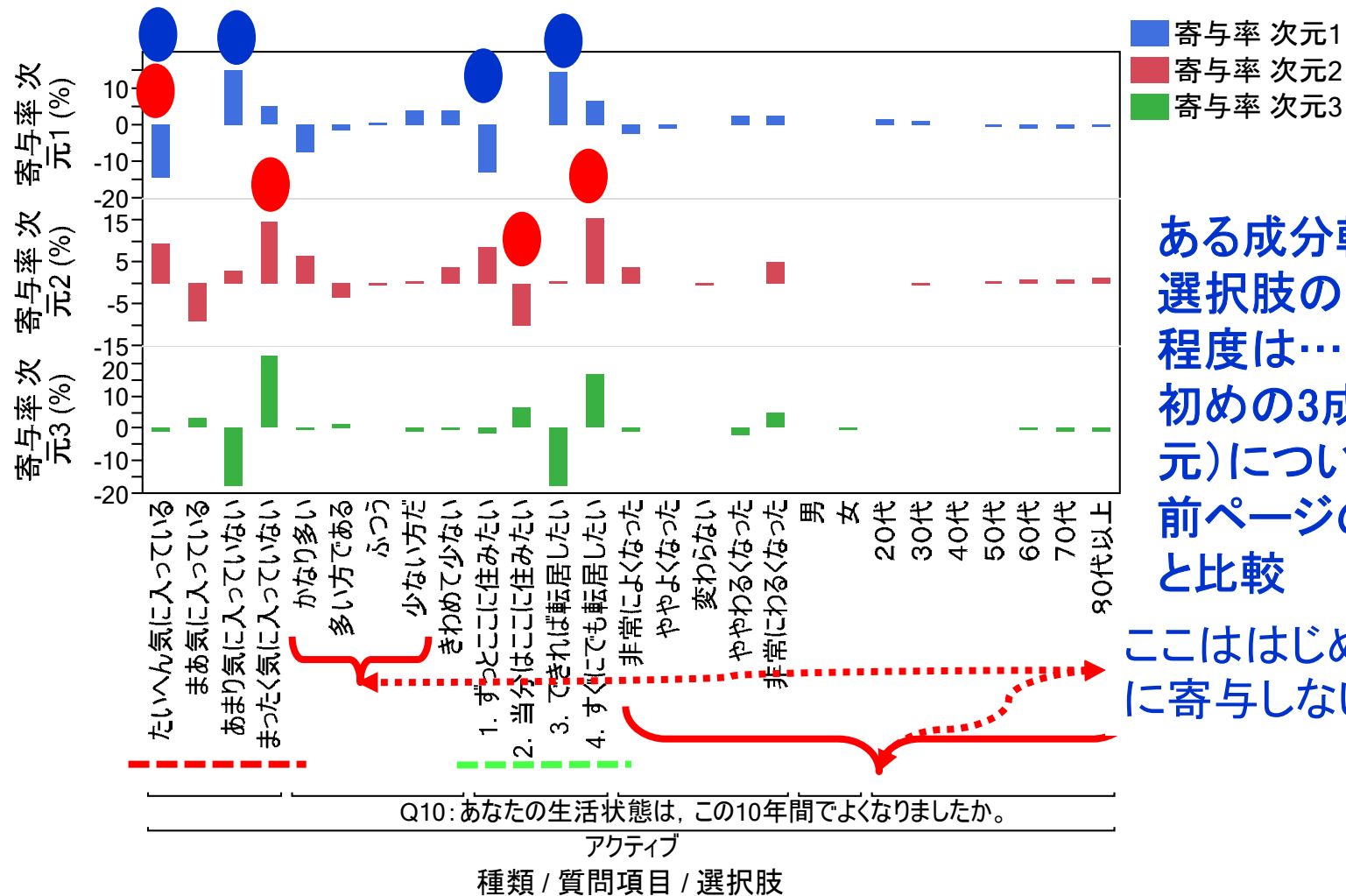
ここで寄与度を参考にする

- MCAにおける寄与度の利用法は、いろいろと研究課題がある。さらに検証が必要である(検証中の課題)。
- 一つの実験的試みとして説明する。
- 絶対寄与度と相対寄与度を、はじめの“3成分”について求め、グラフィカルに表示する。
- 絶対寄与度については、成分スコアの符号で振り分けるようにした(とくに意味は無い、布置図との対応を見易くするため)。
- 成分1(**青**)、成分2(**赤**)、成分3(**緑**)を付けて区別した。
- 相対寄与度は、平方相関ともいうように成分軸からの角度(の大きさ)をみるとよい。配色は同じ。

調査地点を除外し、ほかは同じ, ...



絶対寄与度(はじめの3成分)

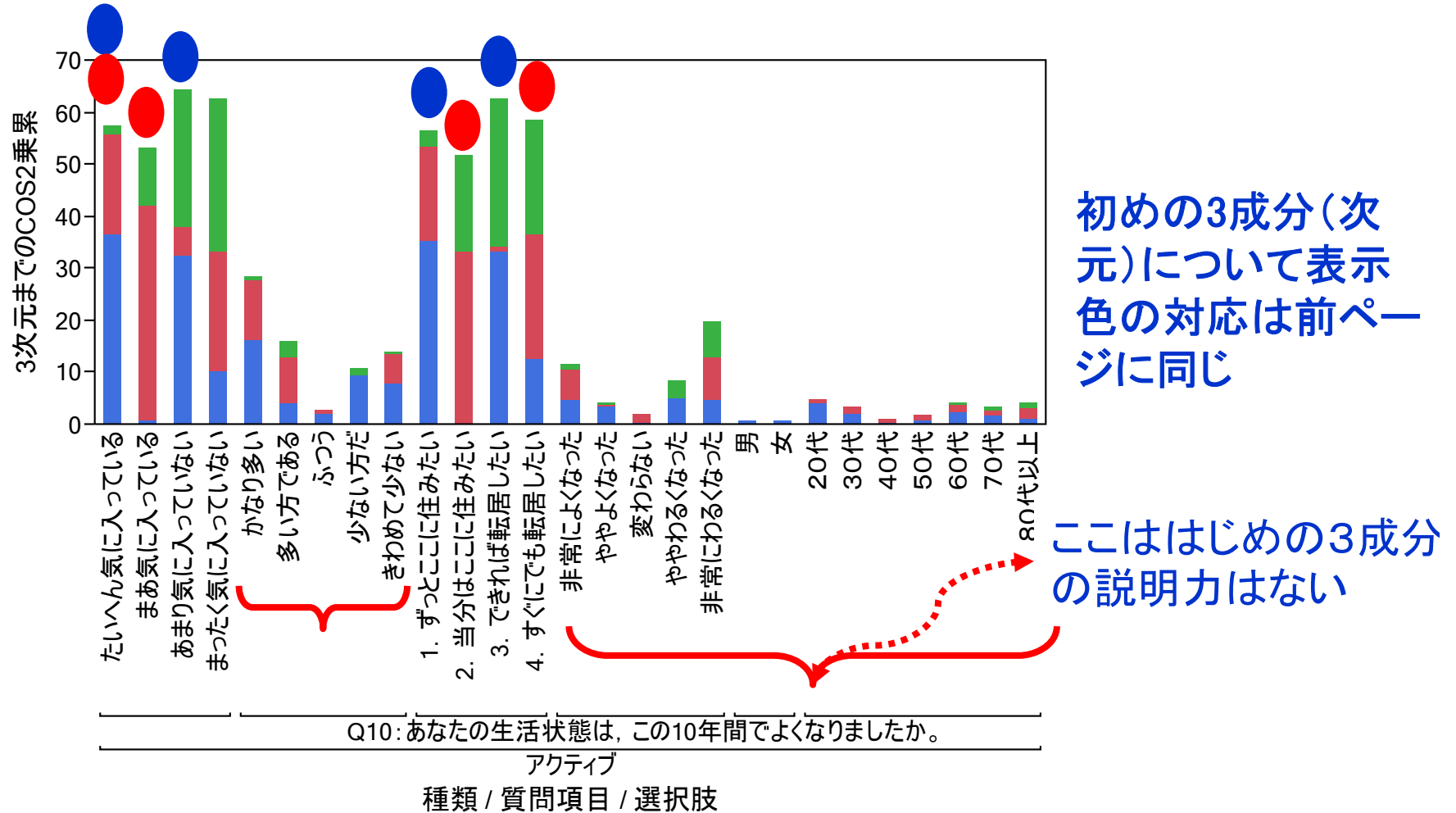


ある成分軸への各
選択肢の寄与の
程度は…
初めの3成分(次
元)について表示
前ページの2成分
と比較

ここははじめの3成分
に寄与しない

質問「生活はこの10年で…」と、「緑が…」が、他と違う傾向
性別、年齢区分が、さほど関与していないのではないか

相対寄与度[平方相関] (はじめの3成分)



各選択肢が, どの成分でよく説明されるか (近似がよいか) をみる. ここでも類似の傾向がみえる.

- これら質問項目には,「調査地点」の影響が大きいようだ.
- 4つの質問項目も,「住むまちが気に入っているか」「ずっと住みたいか」が支配的,これは予想されること.
- 「緑が多いか」「この10年でよくなったか」は,すこし異なる傾向にある.
- 一方,「性別」「年齢区分」はいずれにも関与の程度がほぼ類似し,質問項目との関連が似た傾向にあるようだ.
- では,シナリオを変えて,他の質問項目の組合せとしたらどうだろうか...,と探査する.
- **注意:** 実は,人口統計学的変数と調査地域の回収標本の傾向に”ごくわずかの差違“がある. これはどう影響するか,単項目分析に戻ってみよう. **[回収標本と計画標本との偏り]**

例2:「自治体市民調査」から

- 別の例として, ある自治体で行った小規模の「市民意識調査」の回収データを調べる.
- 調査当時, この自治体には農水省との関連で, ある「農業公園」という施設を運営していた(いまもある).
- この公園の認知度, 利用度ほかを調べることを目的に行った調査.
- 当時は, 市民はこの調査課題に関心なく, 回収率は低かった.
- つまり計画標本に対して偏りがある. 確率標本ではあるが, 回収時点で, 全市民の意見を代表しているわけではない.

(つづき)

- 最近、この市以外の認知度も高くなり来訪者が増えていると聞く.
- 状況が変わって、市民の公園への人気が出てきて、利用者が増えている.
- 再調査するとよさそう. 意見の変化が見えるはず.
- 調査の概要については、演習用データに添付した. スライド(その1)と、テキスト第 I 部の26ページあたりに一部記述. データ表の説明に用いた.

調査情報の再確認

- ある自治体(柏市)で行った市民意識調査の例.
- 市内にある「農業公園についての意識調査」
- 計画標本の大きさ:1,008人(男性493人, 女性515人)
- 回収標本の大きさ:411人(男性178人, 女性233人)
回収率:41% (*)高いとはいえない回収率
- 調査対象者:市内に居住の成人
- 選挙人名簿から2段無作為抽出, 確率標本
- 調査方式:郵送調査, 自記式方式
- 調査期間:平成3年10月(1991年)～平成4年2月(1992年)

調査データの確認

分析に用いる項目

あなたは今の生活環境のなかで日頃どのような過ごし方をしていますか.

- 近くの緑地や公園等をよく散策している.
- 昔からの習慣をよく守っている.
- 神社や, お寺詣りをよくする.
- 自分のなすべき役割は積極的に果している.

ここで, 選択肢はそれぞれ4段階の名義尺度

1. そうではない 2. ややそうではない 3. ややそうだ 4. そうだ
項目の選択肢数は, 「 $4 \times 4 = 16$ 」

人口統計学的変数ほか

性別(2選択肢), 年齢区分(6区分), 性年齢区分(12区分)

年齢区分は「標本抽出枠情報」を利用.

居住地域区分(5区分)[注: 調査課題の公園からの距離区分]

4質問, 性別, 年齢区分(固有値, 寄与率ほか)

- いきなり“全項目”を指定してMCAを実行してみよう. ただし, 属性は「性別」「年齢区分」を用い, 「性年齢区分」は用いていない(こちらを使うとどうなるか).
- 7項目, 延べの選択肢数が29となった.
- データ情報と得られた固有値情報は以下となった.

データ情報

データ名: 市民意識調査(411s)_[MCA]

データテーブルの行数	411
アクティブな調査対象者数	388
アクティブな質問項目数	7
アクティブな選択肢数	29
分析に使われたアクティブな選択肢数	29
追加の調査項目者数	0
追加の質問項目数	0
追加の選択肢数	0

固有値の和

固有値の和	3.14286
固有値の2乗	0.51225
Benzécri和	0.05722
Greenacre和	0.07382
非対角の2乗	0.06327
対角の2乗	0.44898

2乗和において対角ブロック部分が占める割合:
87.65%

固有値, 寄与率, 累積寄与率





































- インジケータ行列(A), バート表(B)から得た情報を示した.
- Benzécri, Greenacreの調整済み固有値ほかは次ページ.

次元	固有値		割合(%)	累積(%)	固有値 2乗		割合(%)	累積(%)
1	0.295		9.38	9.38	0.087		16.95	16.95
2	0.246		7.81	17.19	0.060		11.77	28.72
3	0.212		6.74	23.93	0.045		8.75	37.48
4	0.185		5.89	29.82	0.034		6.70	44.17
5	0.175		5.58	35.40	0.031		6.01	50.18
6	0.164		5.22	40.63	0.027		5.26	55.44
7	0.158		5.03	45.66	0.025		4.89	60.33
8	0.154		4.89	50.55	0.024		4.60	64.93
9	0.149		4.73	55.28	0.022		4.32	69.25
10	0.138		4.39	59.67	0.019		3.72	72.98
11	0.134		4.27	63.94	0.018		3.52	76.49
12	0.132		4.20	68.15	0.017		3.40	79.90
13	0.125	λ_k^A の確認	4.13	72.28	0.016	λ_k^B の確認	2.96	82.86
14	0.120		3.96	76.24	0.015		2.79	85.65
15	0.109		3.48	79.72	0.012		2.34	88.00
16	0.108		3.44	83.16	0.012		2.28	90.28
17	0.104		3.30	86.46	0.011		2.10	92.38
18	0.103		3.26	89.72	0.011		2.05	94.43
19	0.095		3.04	92.76	0.009		1.78	96.21
20	0.086		2.75	95.51	0.007		1.46	97.67
21	0.079		2.51	98.02	0.006		1.21	98.88
22	0.071		2.25	100.00	0.005		0.98	100.00

$$\frac{1}{M} = \frac{1}{7} = 0.1429$$

Benzécri, Greenacreの調整済み固有値

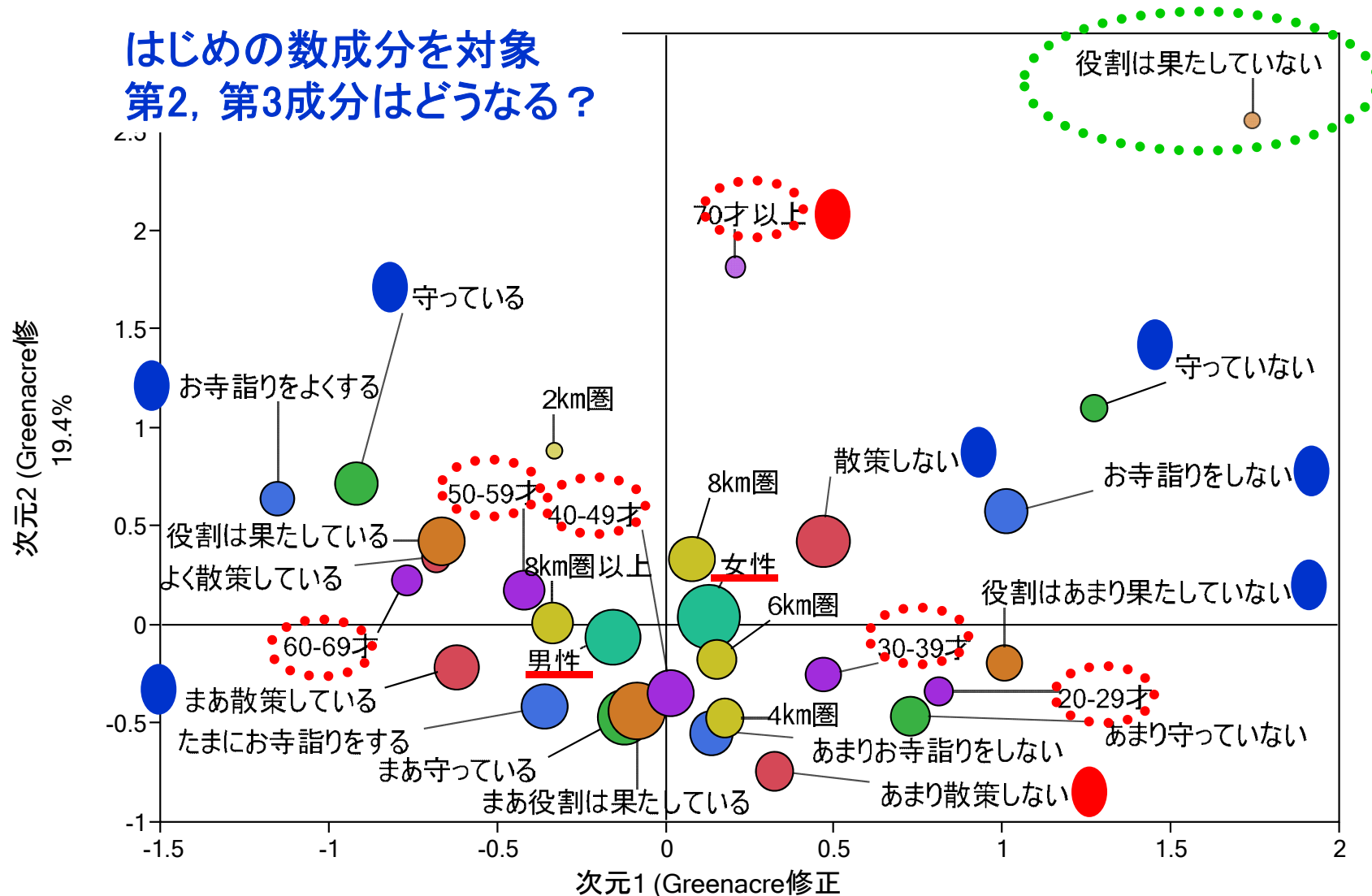
- 先頭から2, 3成分を採用すればよさそう.

調整固有値(2乗)	Benzécri 割合(%)	Benzécri	Benzécri 累積(%)	Benzécri	Greenacre 割合(%)	Greenacre	Greenacre 累積(%)	Greenacre
0.03	54.85		54.85		42.51		42.51	
0.01	25.08		79.93		19.44		61.96	
0.01	11.28		91.21		8.74		70.70	
0.00	4.28		95.48		3.31		74.01	
0.00	2.52		98.01		1.95		75.97	
0.00	1.08		99.08		0.84		76.81	
0.00	0.56		99.65		0.44		77.24	
0.00	0.27		99.92		0.21		77.45	
0.00	0.08		100.00		0.06		77.52	
.	
.	
固有値の和					.		.	
固有値の和					.		.	
固有値の2乗和					.		.	
Benzécri和					.		.	
Greenacre和					.		.	
非対角の2乗和					.		.	
対角の2乗和					.		.	

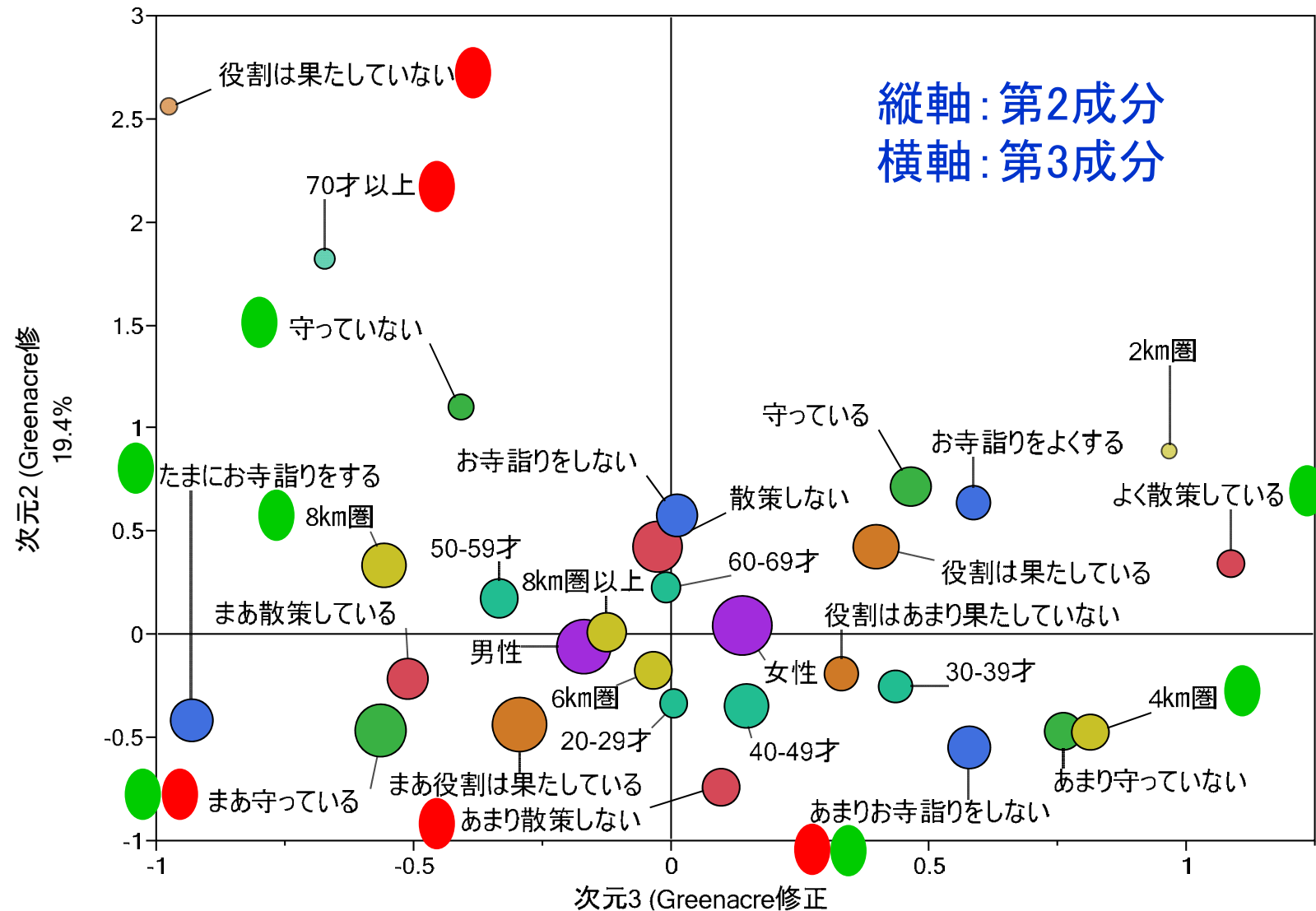
2乗和において対角ブロック部分が占める割合 : 87.65%

項目・選択肢の成分スコア（布置図）

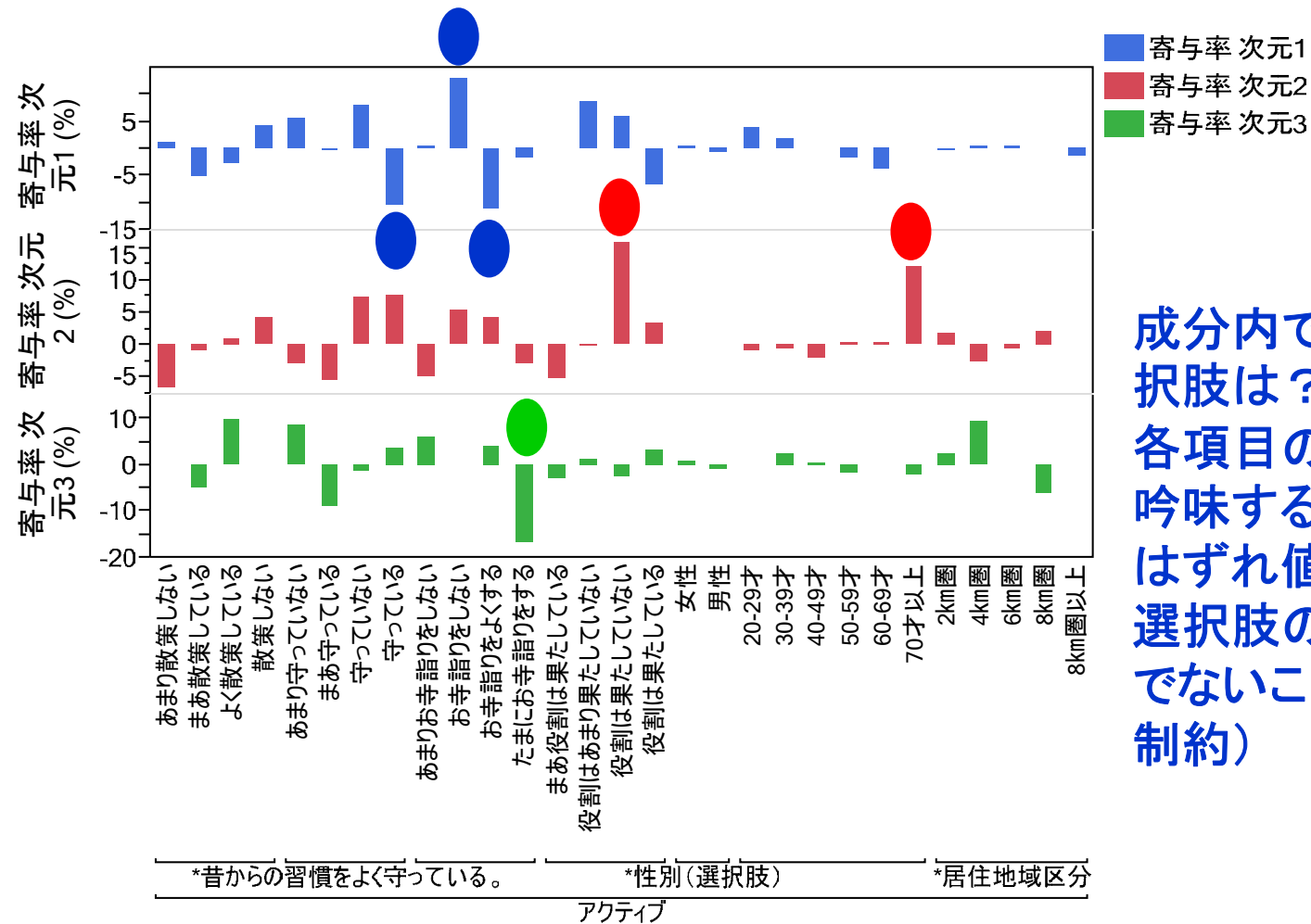
はじめの数成分を対象
第2, 第3成分はどうなる？



第2, 第3成分で観察すると, ...

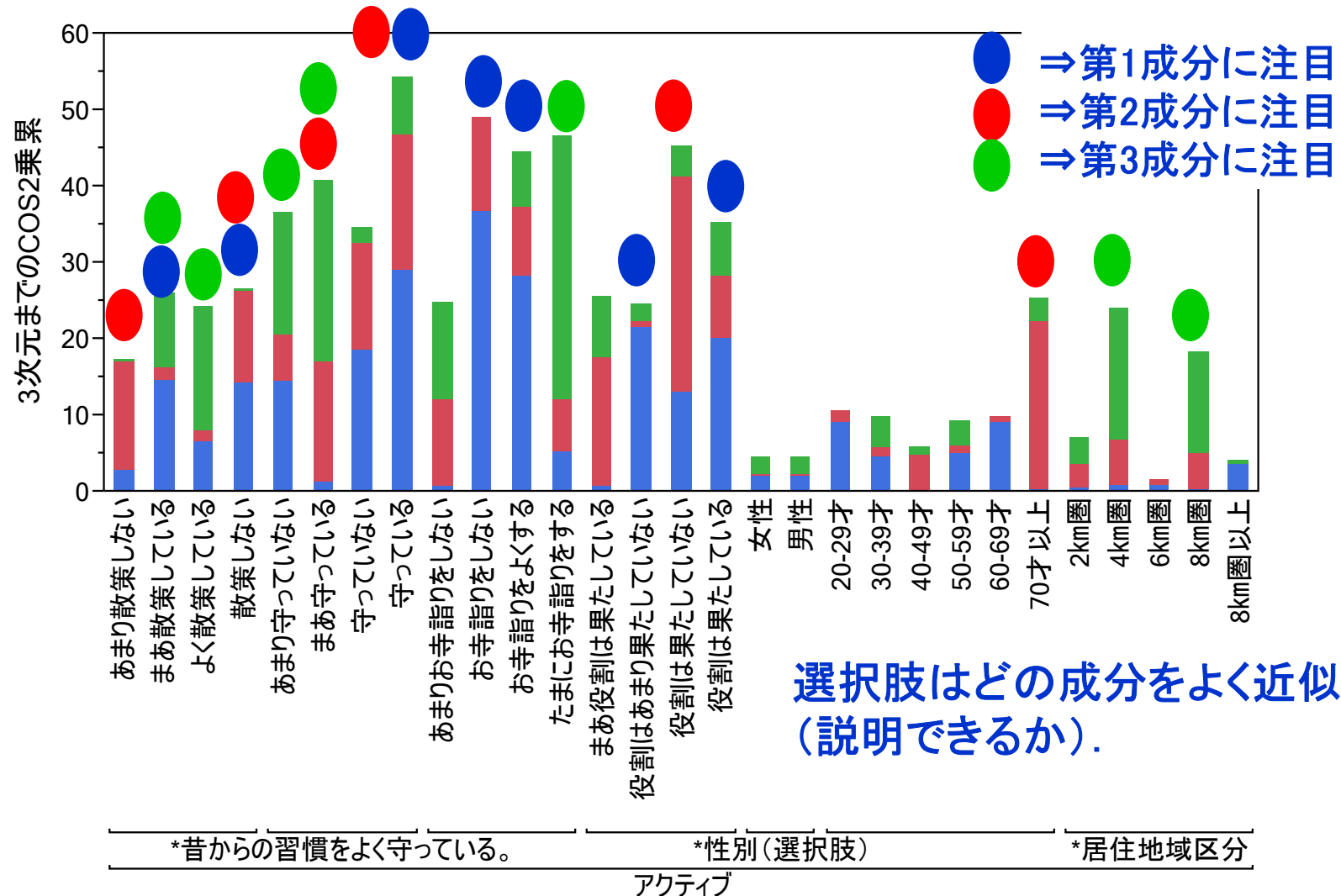


絶対寄与度

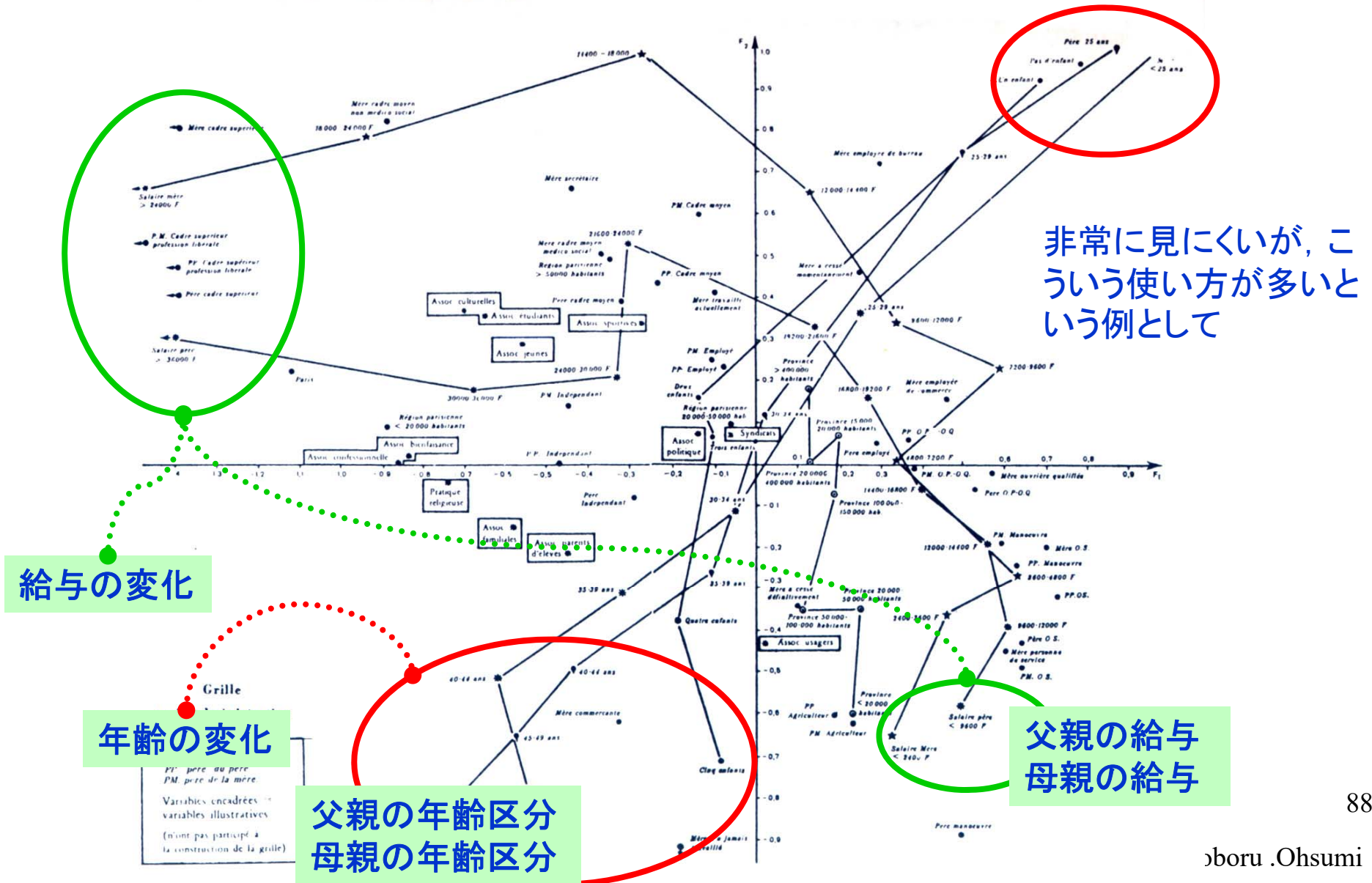


成分内で高く寄与する選択肢は？
各項目の選択肢の傾向を吟味する
はずれ値的な選択肢は？
選択肢の並び順が元の順でないことに注意(ソフトの制約)

相対寄与度



Volleの本“Analyse des Données”から



P. Bourdieu(ピエール・ブルデュー)の研究

- ピエール・ブルデュー(Pierre Bourdieu, 1930年8月1日～ 2002年1月23日)は, フランスの著名な社会学者.
- 著書「ディスタンクション」(disitinction)の中で, 対応分析・多重対応分析を多用している(ことでも, 有名).
- INSEE(フランス統計局)などとの共同研究で, 多数の社会調査を行っている. 実証＋思想
- ディスタンクションとは「区別・差異化・卓越化・上品さ...」など多様な意味(含意)がある.

“La Distinction, Critique sociale du judgement” (1979)

英語翻訳版あり(1987)

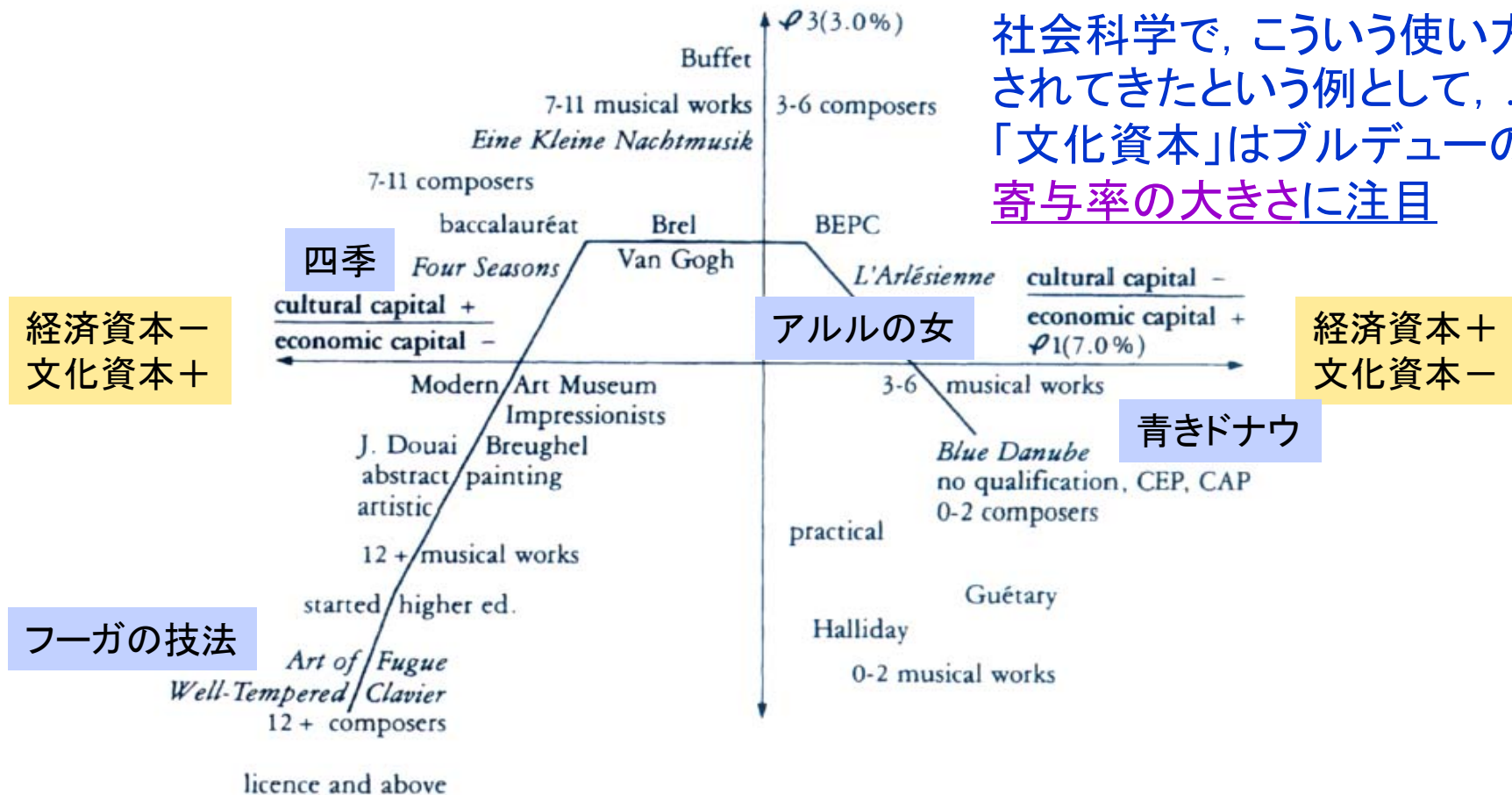
“Disitinction: A Social Critique of the Judgement of Taste”

「ディスタンクション, 社会的判断力批判」, 藤原書店刊.

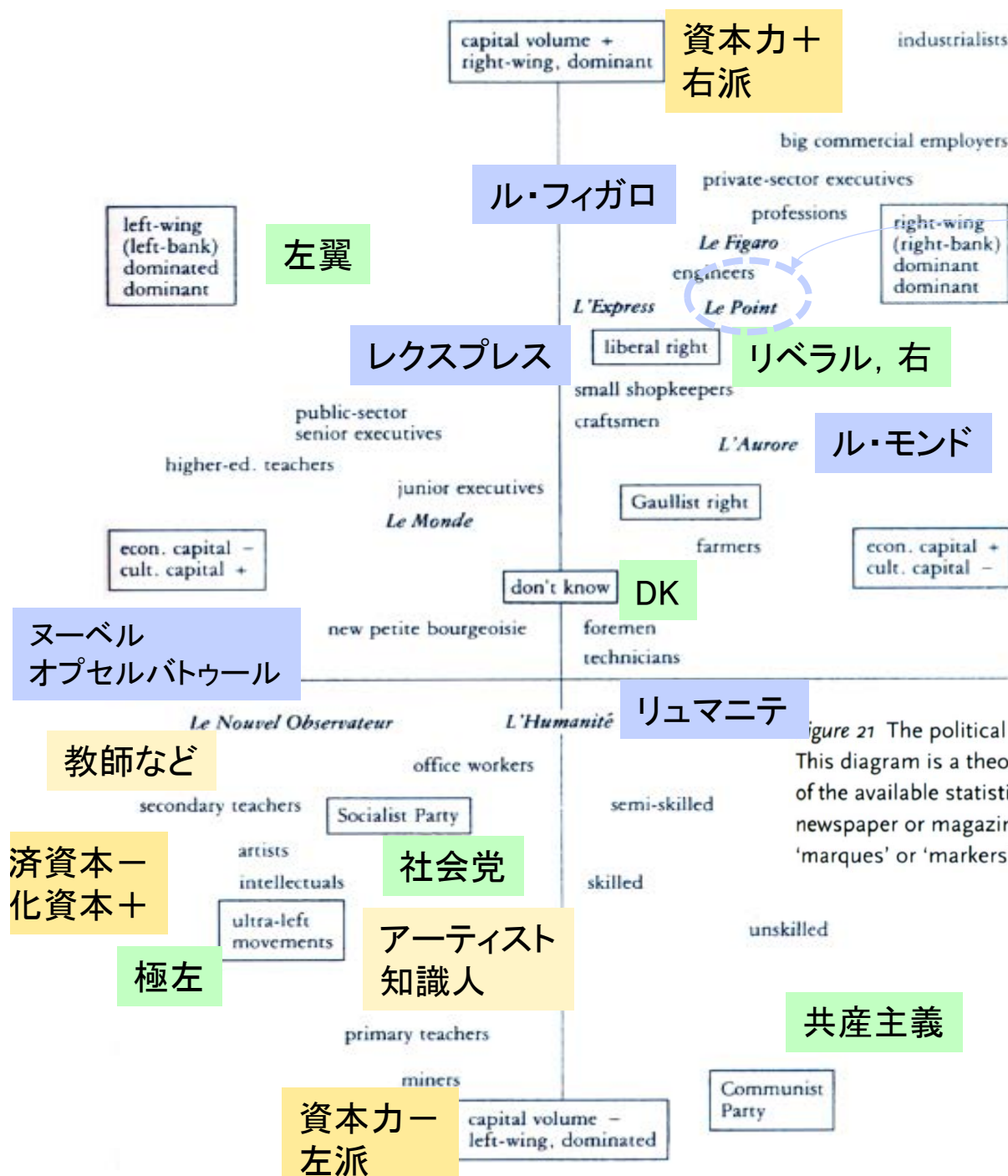
- フランスの諸階級の文化的嗜好に関する研究, 政治の考え方, 趣味嗜好などを通じて, 社会階級と文化資本, 階級文化と教育, ...と多彩な研究.
- たとえば, 教育と社会階級について分析. 裕福な家庭の子が進学で有利, 文化資本(上品で正統とされる文化や教養や習慣等)の保有率が高い学生ほど高学歴である等々, を統計分析的に示し議論した.

- マルクス主義, さらには新自由主義やグローバリゼーションに批判的であった(とされる).
- 「ディスタクシオン」の中で無数に用いられた対応分析の引用例.
- 何回もの調査を行い, 仮説発見“と”検証“を繰り返す.

美術, 嗜好, ... (文化資本と経済資本)



社会科学で, こういう使い方がなされてきたという例として, ...
「文化資本」はブルデューの主張 寄与率の大きさに注目



資本力+
右派

ル・ポアン

ル・フィガロ

左翼

右翼

レクスプレス

リベラル, 右

購読紙・誌,
政治的立場, ...

ル・モンド

DK

経済資本+
文化資本-

ヌーベル
オプセルバトゥール

リュマニテ

教師など

済資本-
化資本+

極左

社会党

アーティスト
知識人

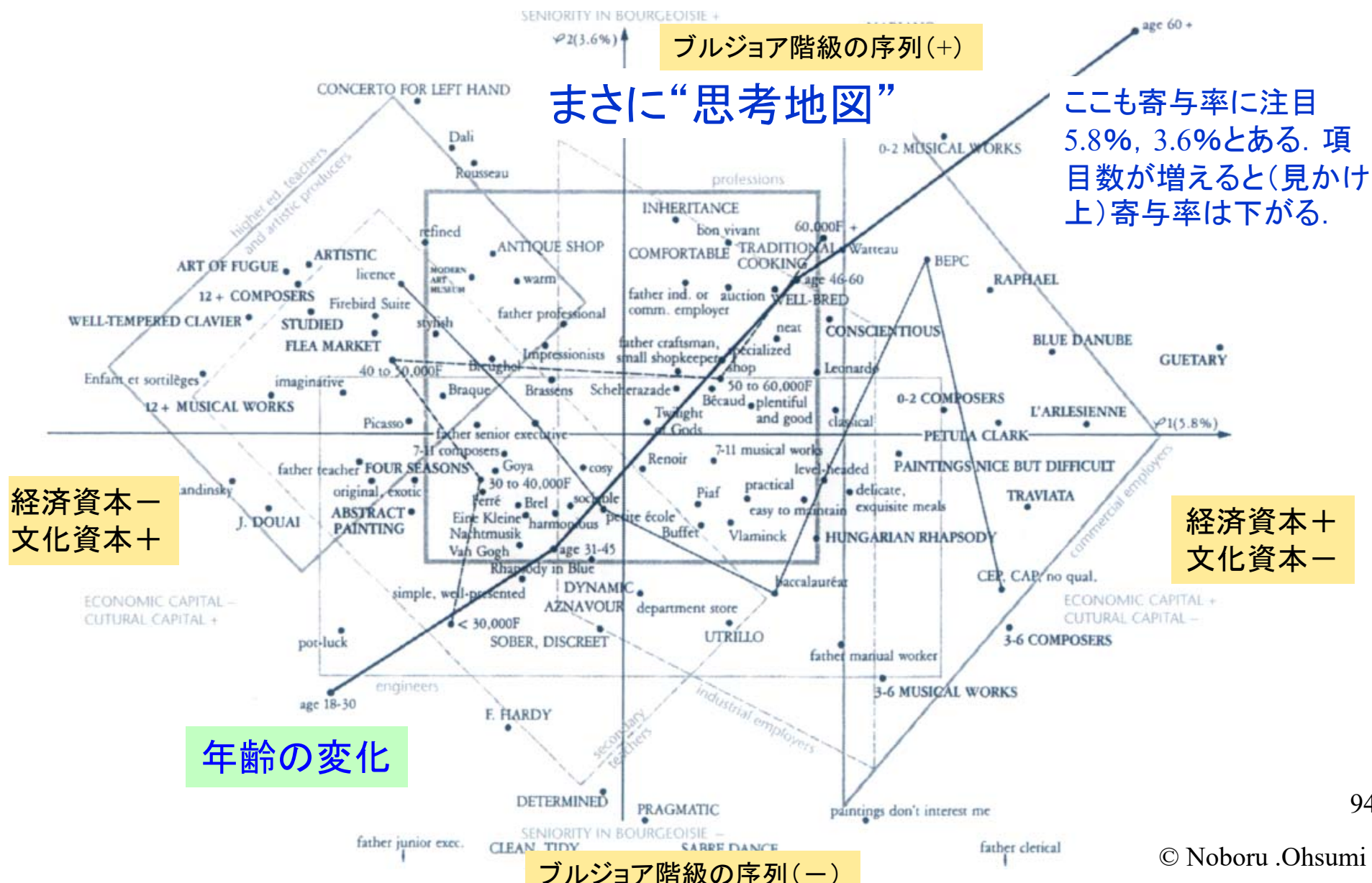
共産主義

資本力-
左派

Figure 21 The political space.

This diagram is a theoretical schema constructed on the basis of a close reading of the available statistics (and various analyses of correspondences). The only newspaper or magazine titles indicated are those which function as political 'marques' or 'markers'.

“Distinction” (ディスタクション), 5章から



おわりに常識的な指摘をすこし, ...

- MCAスクリプト(JMPスクリプト)を用いて, 構造が比較的簡単なデータ表を分析する例を示した.
- 現実の調査データでは, 項目数が非常に多く, ときには選択肢数も多くなる傾向にある.
- 調査票・質問文の設計時の検討が重要であること. やたらに選択肢を増やさない, ワーディングに注意する, ニュートラルとなる選択肢に注意する, 等々.
- とくにウェブ調査では注意する.
- 予備調査・パイロット調査を行って, 調査票のできればえを調べる.
- 調査関係者(委託者, 実施者)だけのチェックでもよい.

(つづき)

- 過去の調査から、継続的に利用できる“再現性”のある、あるいは“安定した傾向”を示す質問文をためておく.
- “外部情報源”の利用を検討する. それを調査にどう組み入れるかを事前に設計する.
- とくに, “非標本誤差”の回避への対応策. 無回答バイアスや無回答の低減を念頭に対処.
- 調査方法論. 調査技法の知識を反映させること.
- “調査方式・データ収集方式”の周到な準備が決め手.

テキスト, 参考文献を参照.

大隅(監訳)「調査法ハンドブック」(朝倉書店)など

Tourangeau他(2013): “The Science of Web Surveys”(翻訳刊行予定)