セキュリティ技術者のための カテゴリカルデータの 統計分析法入門 ver1.1

ITリスク研究会報告
2023年7月22日（土）
津田塾大学 数学・計算機科学研究所
国立情報通信研究機構（NICT）サイバートレーニング研究室
藤本一男
kazuo.fujimoto2007@gmail.com

案内の口上書き

- セキュリティに関連して調査などを行うと数値としては取得されないデータ（カテゴリカル・データ）の扱いが必須となります。
- たとえば、５、４、３、２、１というコードが回答「とても当てはまる」「当てはまる」「どちらでもない」「あまり当てはまらない」「全く当てはまらない」に対応して付けられているとします。
- このような設問が複数用意されている時に、その設問ごとの平均値、分散を計算して比較することはデータ構造を破壊して分析しているのです。
- こうした場合に適用できる手法として「**対応分析**」という手法があります。
- 今回の報告では、対応分析がどのようなロジックでカテゴリカル・データを「数量化」し、統計処理するのかをいくつかの事例を交えてご紹介します。

研究テーマ：「対応分析」

- 2004年ごろ？Rと出会う。Ver1.9x？
- 社会調査実習の指導で使う。
- 「対応分析」との出会い Applied Correspondence Analysis の翻訳本の「解説編」でRで検算を書く。『対応分析入門』2015年
- 2020年11月翻訳『対応分析の理論と実践』

研究テーマ（その2）

- 科研費「データの幾何学的構造に注目したカテゴリカル・データの研究」★これが本命
  - KAKENでの説明 https://nrid.nii.ac.jp/nrid/1000040348090/
  - 「対応分析」ってなんですか、というコラム
    - 作新学院大学の図書館ニュースレター：https://bit.ly/2XyorN2

- 近似された運動強度として心拍測定/鼻呼吸継続度を元にLT（乳酸閾値 lactate threshold）直前のペース走をモニタする方法の研究
  - 趣味のランナーです。もう歳なので、無理せずノンビリ、でも軽快に！をモットーに走ってます。

今日のお話の構成

- カテゴリカルデータの扱いは、なかなか大変な状況にあります。
- 広義の情報システムの定義を考えたとき、調査データでカテゴリカルデータの扱いは不可欠となります。

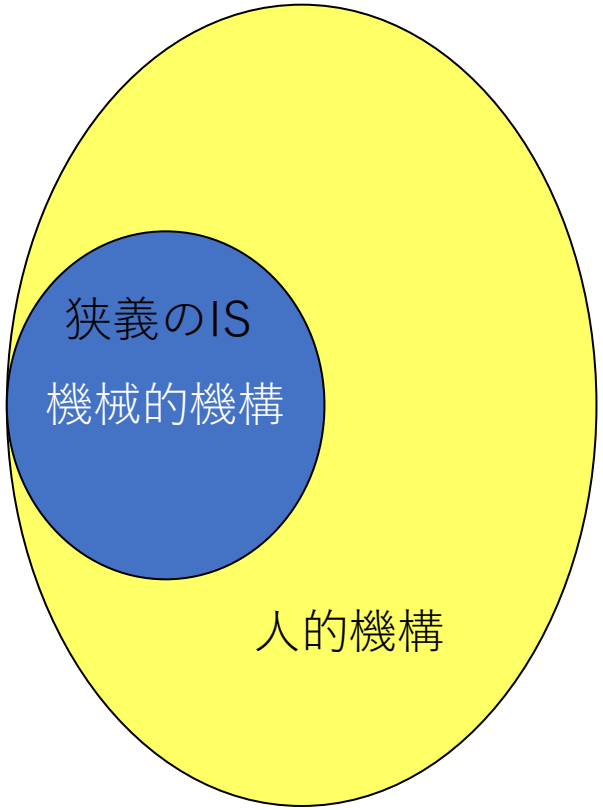
- そこで、可能な限りデータ構造を破壊せずにカテゴリカルデータを分析する手法（統計的手法というよりもデータ処理観）としての**対応分析**（Correspondenece Analysis）という手法をご紹介します。

はじめに：「セキュリティ技術」と調査データをめぐったメモ

- 情報システム（IS）の概念と人的機構
  - 狭義のIS
  - 広義のIS
- 浦昭二先生たちのテキスト『情報システム学へのいざない』初版1998年、第２版2008年
- セキュリティを問題にするときには、この**広義のIS**の視点が不可欠。
  - 利用者アンケート、管理者アンケートなど
  - 社会調査のデータは、大半がカテゴリカルデータ。

情報システムとはなにか

- 情報システムとは、組織体（または社会）の活動に必要な情報の収集・処理・伝達・利用に関わる仕組みである。
- 広義には**人的機構**と**機械的機構**とからなる。
- コンピューターを中心とした機械的機構を重視した時、**狭義**の情報システムと呼ぶ。
- しかし、このときそれが置かれる組織の活動となじみのとれているものでなければならない。
- ［浦・他1998:p40, 2008:p53］



広義のIS

コード化の例

- よくあるコード化、５、４、３、２、１
  - 整数尺度、リッカート尺度
- そのまま数量データとして加算して合計点、平均や分散を計算

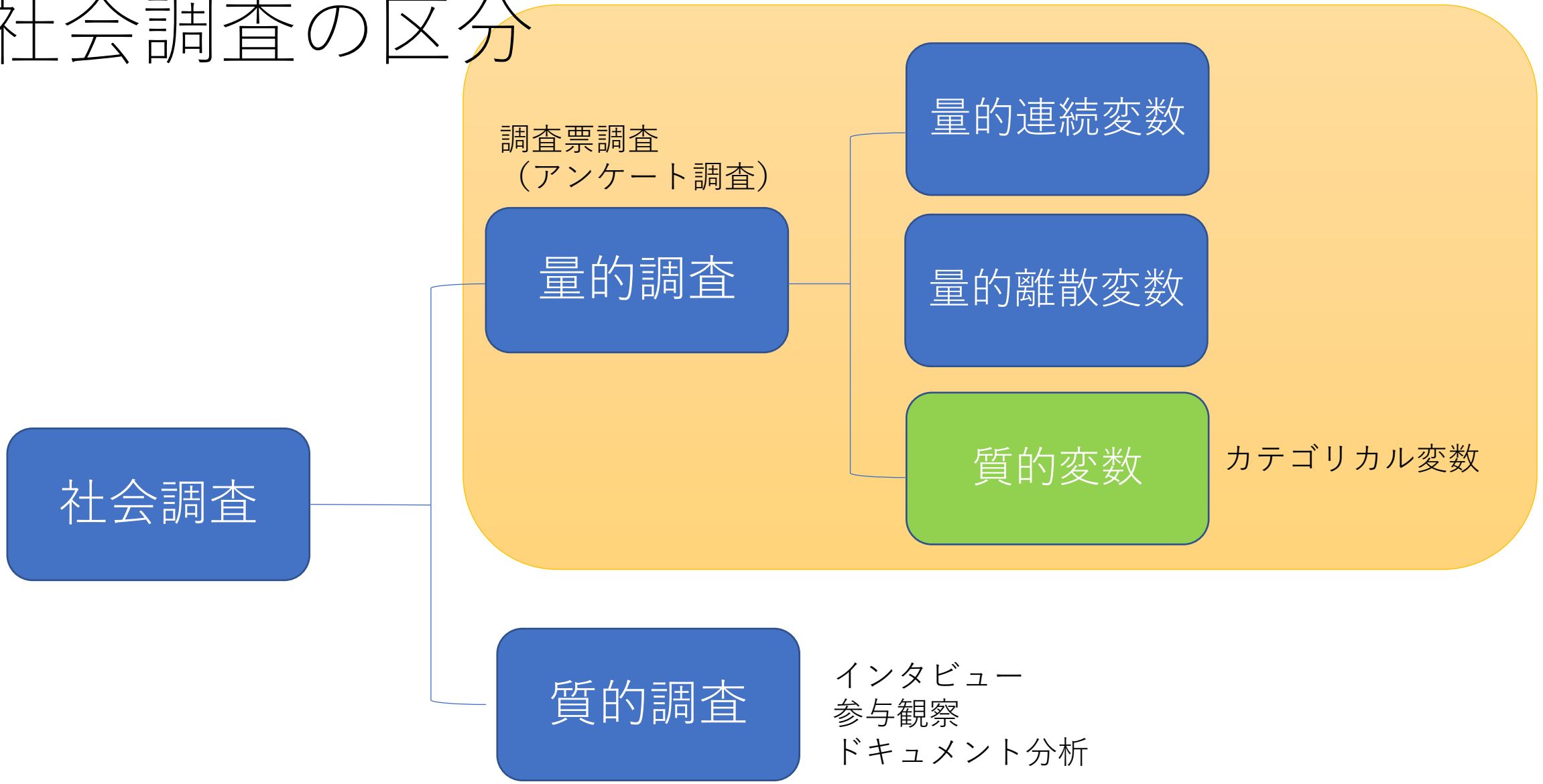
- それ、大丈夫ですか。
  - 前提にできますか？
    - 設問ごとのWeight
    - 等間隔（整数）？
    - リニア？

A. あなたの仕事についてうかがいます。最もあてはまるものに○を付けてください。

		そうだ	まあそうだ	ややちがう	ちがう
1.	非常にたくさんの仕事をしなければならない	1	2	3	4
2.	時間内に仕事が処理しきれない	1	2	3	4
3.	一生懸命働かなければならない	1	2	3	4
8.	自分のペースで仕事ができる	1	2	3	4
9.	自分で仕事の順番・やり方を決めることができる	1	2	3	4
10.	職場の仕事の方針に自分の意見を反映できる	1	2	3	4

厚労省：職業性ストレス簡易調査票 （簡略版23項目） より抜粋
https://stresscheck.mhlw.go.jp/download/material/sc23.pdf

社会調査の区分


社会調査
量的調査
調査票調査
（アンケート調査）
量的連続変数
量的離散変数
質的変数
カテゴリカル変数
質的調査
インタビュー
参与観察
ドキュメント分析

データ構造を破壊しない手法？

- 数理統計学は、連続量データと確率分布を要素として発展してきた。
  - 離散データも、連続データで近似。
  - カテゴリカルデータも？

- 性別、出身地、専門、など、そのままでは数値にできないデータが溢れている。
- 社会調査データは、こうしたデータのオンパレード
  - 統計処理するには「数量化」という処理が必要。
  - もう一つ、「多次元データ」としてあつかう、という視点も。

- 「対応分析」は、こうした数値ではないデータを、多次元性を維持し、「数量化」します。

対応分析はどのような手法か

Correspondence Analysis

名称：CAとMCA

- CA（Correspondence Analysis）は、２変数（クロス表）データの分析。
- MCA（Mulitiple Corresponcdence Analysis）は、３変数以上のデータ（調査集計表のように、個体x変数）の分析

- ★どちらも、行と列の２変量データの分析！

CA、MCAの仕組み

- 行方向/列方向に、プロファイル（比率）ベクトルをつくり、そうやって定義される「点」の空間を考え、次元縮減する。
- 数理的なコア
  - 同時確率行列をもとに、標準化残差（期待値との差）の特異値分解によって、次元を縮減する。主成分分析（PCA）でやるのと同じ。
  - 行空間を列空間が生成され、
  - 各点がPlotされます。

CA、MCAの応用

- CA/MCAは、行空間（個体空間）と列空間（変数空間）を生成しますが、その空間の座標軸が有している分散（情報量）は、同じになります。（距離を$\chi 2$距離で評価するため。）
- また、その空間の点の位置は、相互に浸透している（対応している）関係にあります。
- そこから、他方からもう一方に、空間には影響をあたえずに、点を射影する、という方法が可能になります。
- この特性をいかして、空間生成に寄与する変数と射影し空間を説明する変数に区分（構造化モデリング）する、という手法も開発されています。

対応分析の実際

まず事例

- 対応分析の仕組みを聞いてもらうためにまず事例のリスト
- シンプルCA
  - 「職種と余暇の過ごし方」『対応分析入門』の第1章、第９章で使われているデータ。
- 多重対応分析MCA
  - SSM2005から取得した性別役割意識に関する調査データの分析
  - マンガ「因子分析」で主成分分析のサンプルとして使われているラーメン点評価のデータを、カテゴリカルデータとして再分析した例。
  - 某大学の学生生活満足度調査の再分析。

実例１シンプルCA（２変数）

- クロス表をどのように表示するか
- 行分析と列分析
- 対称マップによる同時表示

データ：「職種と余暇の過ごし方」

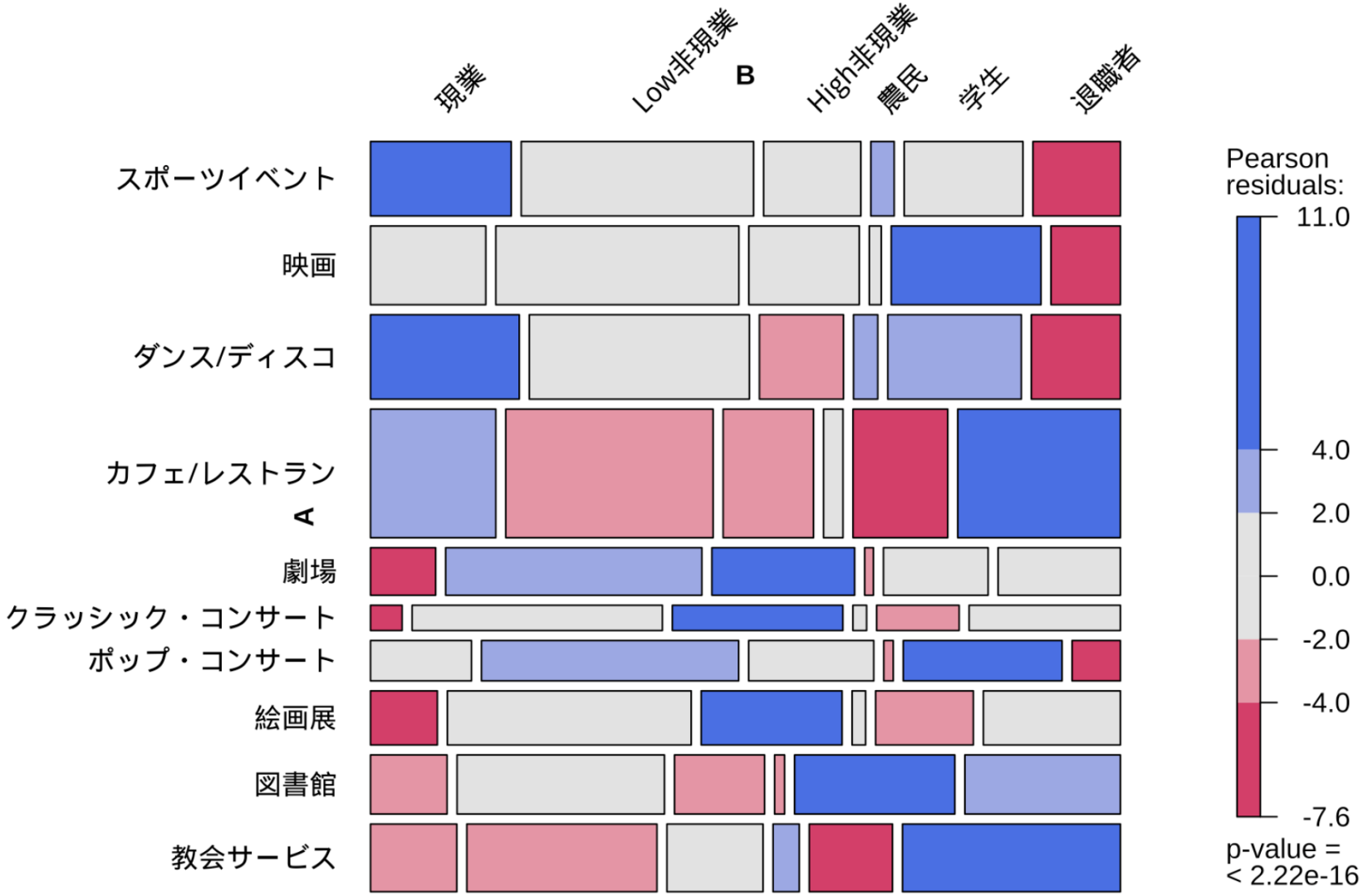
- 行：余暇の過ごし方（10）
- 列：職種（6）
- 10x6 行列

- このデータから
  読み取りたいこと：
  - **職種**と**余暇の過ごし方**の傾向

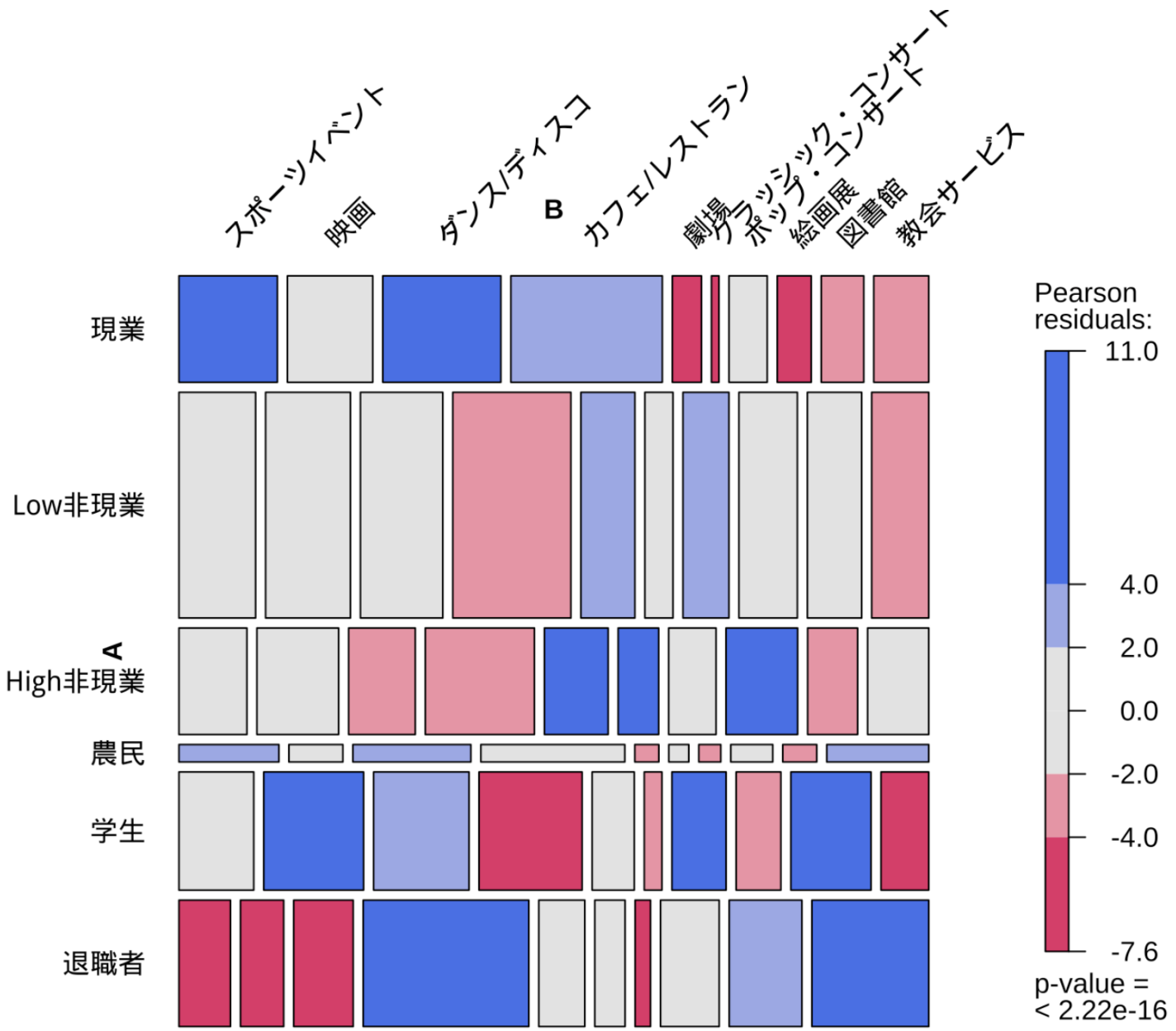
	A	B	C	D	E	F	G
1		職種					
2	余暇	現業	Low非現業	High非現業	農民	学生	退職者
3	スポーツイベント	301	497	208	50	254	187
4	映画	261	550	250	27	339	157
5	ダンス/ディスコ	361	534	204	59	324	216
6	カフェ/レストラン	463	766	334	72	350	601
7	劇場	89	350	195	12	143	167
8	クラッシック・コンサート	23	182	124	10	60	110
9	ポップ・コンサート	117	298	145	11	184	56
10	絵画展	104	379	219	21	152	213
11	図書館	130	352	153	17	272	264
12	教会サービス	168	370	187	51	162	424

行分析

- mosaic plot ：帯棒グラフの帯幅にその帯度数に対応した高さを与えたもの



列分析


スポーツイベント
映画
ダンス/ディスコ
B
カフェ/レストラン
劇場
クラッシック・コンサート
ポップ・コンサート
絵画展
図書館
教会サービス
現業
Low非現業
A
High非現業
農民
学生
退職者
Pearson residuals:
11.0
4.0
2.0
0.0
-2.0
-4.0
-7.6
p-value = < 2.22e-16

ごちゃごちゃしているが…

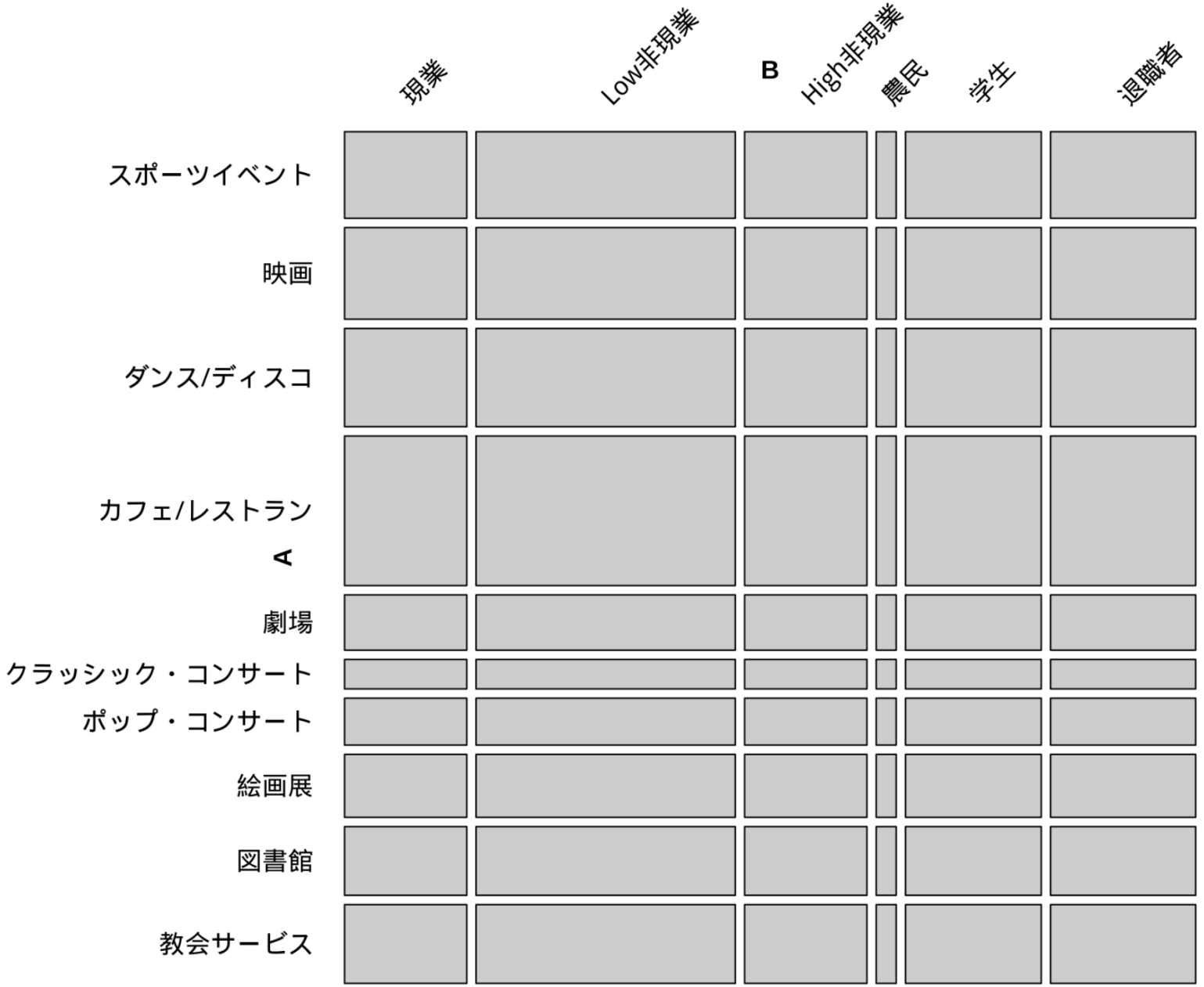
- 「職種」と「余暇の過ごし方」に傾向はない：という状況を考える。
- 「残差」
  - カイ二乗検定でいう期待値状態。二つの変数の間には傾向なし！
  - この状態から各セルがどれだけ離れているのかを、残差（ピアソンの標準化残差）として評価したものが、色付きの部分。

- －2～＋２　：期待値と大差なし
- －4～２、２～４：そこそこ差がある
- －４、４より隔たっている。大いに差がある。

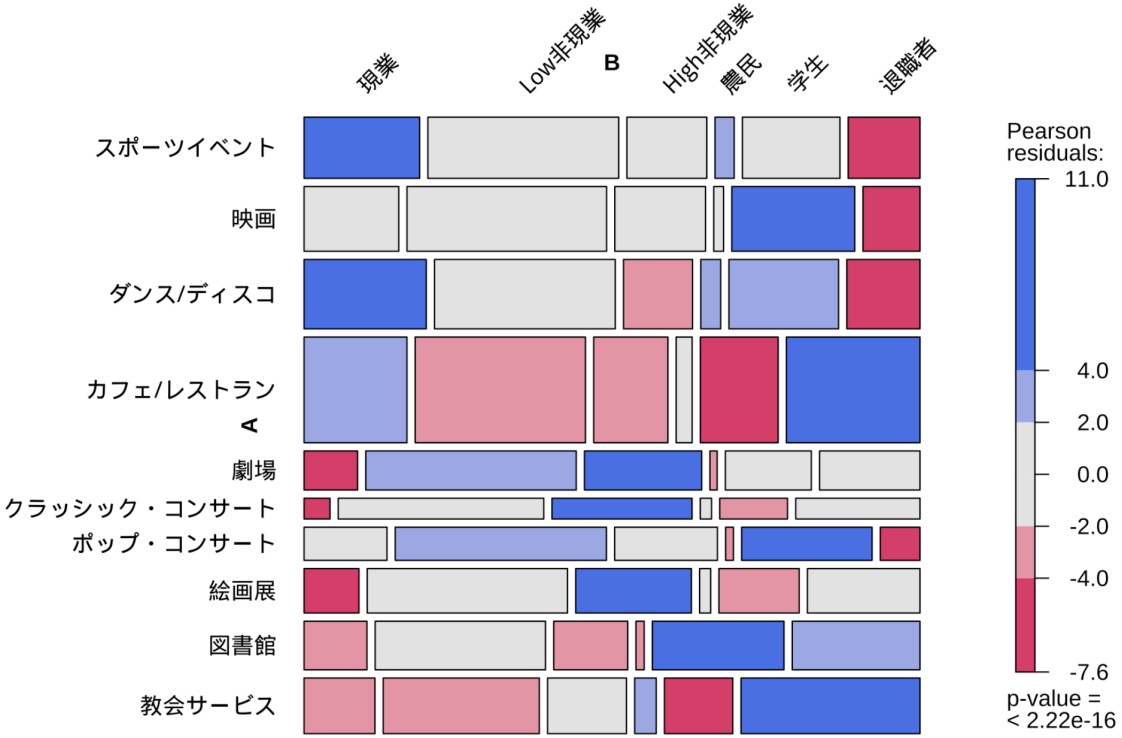
期待値状態

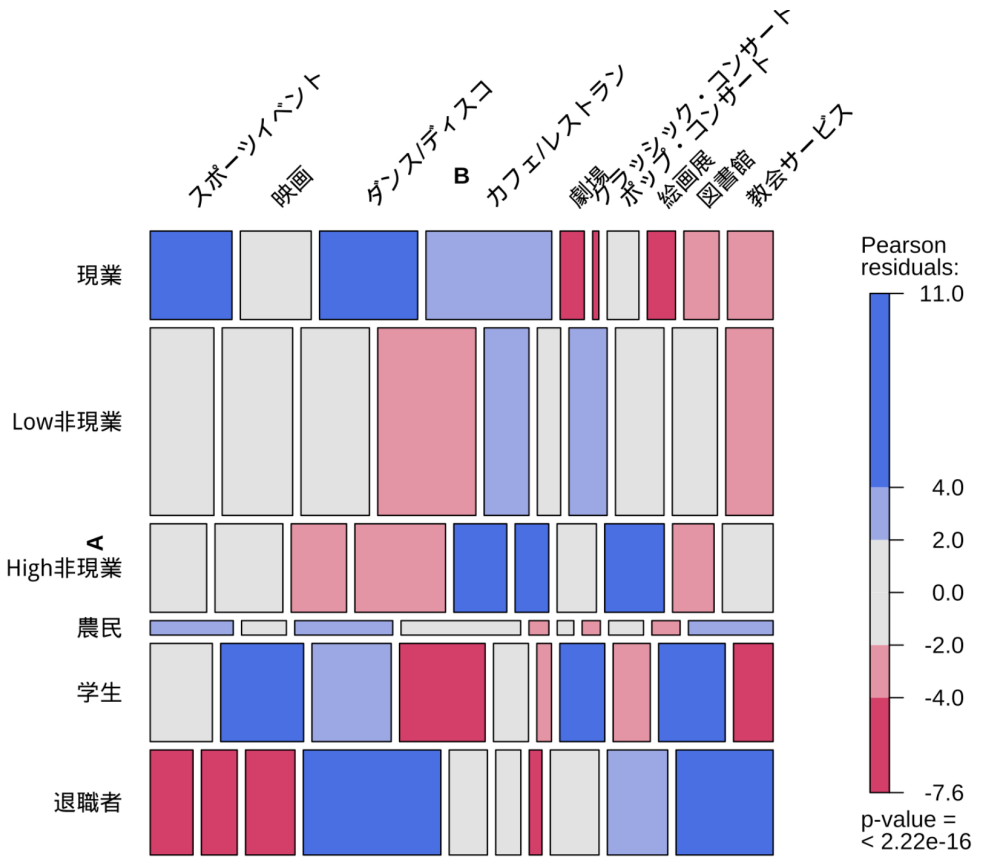
- 標準化Pearson 残差は、

残差＝（セルの度数-期待値）を、行周辺度数、列周辺度数をもちいて、標準化したもの。



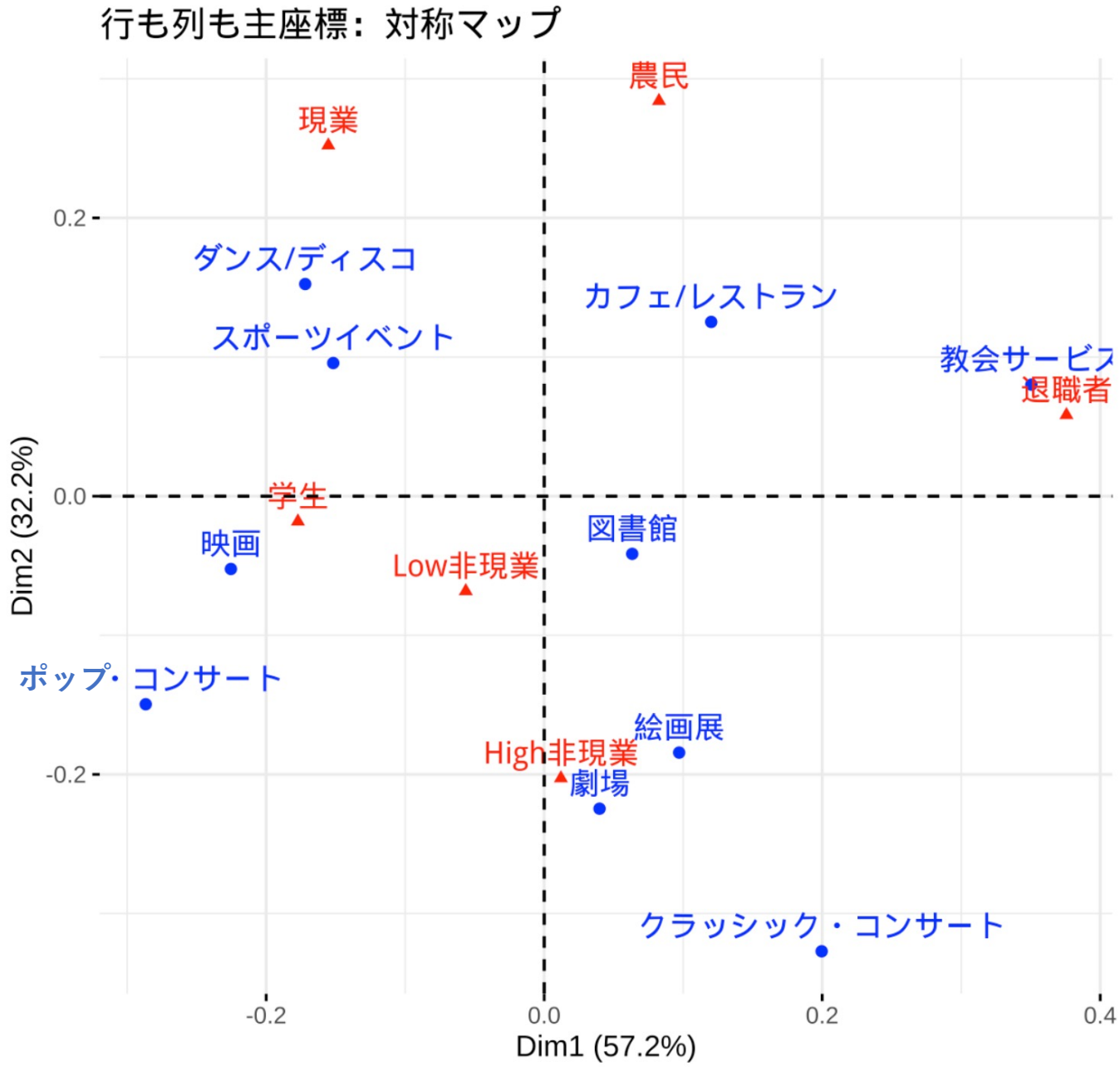
あらためて


現業
Low非現業
B
High非現業
農民
学生
退職者
スポーツイベント
映画
ダンス/ディスコ
カフェ/レストラン
A
劇場
クラッシック・コンサート
ポップ・コンサート
絵画展
図書館
教会サービス
Pearson residuals:
11.0
4.0
2.0
0.0
-2.0
-4.0
-7.6
p-value = < 2.22e-16


スポーツイベント
映画
ダンス/ディスコ
B
カフェ/レストラン
劇場
クラッシック・コンサート
ポップ・コンサート
絵画展
図書館
教会サービス
現業
Low非現業
A
High非現業
農民
学生
退職者
Pearson residuals:
11.0
4.0
2.0
0.0
-2.0
-4.0
-7.6
p-value = < 2.22e-16

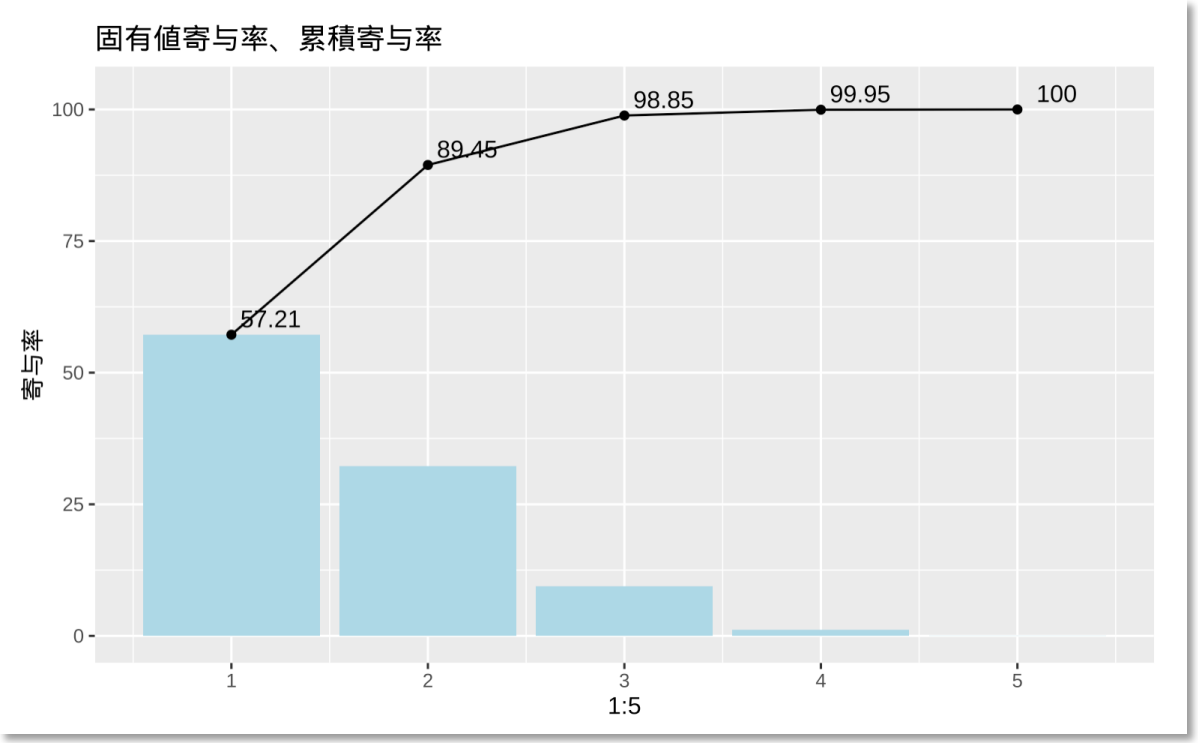
このデータを対応分析します

- .d にデータをセットして。次の一行
- res.CA <- CA(.d)



マップの見方（1）

- 軸の寄与率
  - もともと、10x6行列（6次元）空間のデータを、特異値分解をつかって、次元縮減している。
  - その軸の寄与率でデータ全体の情報（分散）が表現されているかを確認できる。
- 原点は、全体の平均位置。
- 似たものは近くに、異なるものは、遠くに位置する。



固有値 <dbl>	寄与率 <dbl>	累積寄与率 <dbl>
0.04	57.21	57.21
0.02	32.24	89.45
0.01	9.39	98.85
0.00	1.10	99.95
0.00	0.05	100.00

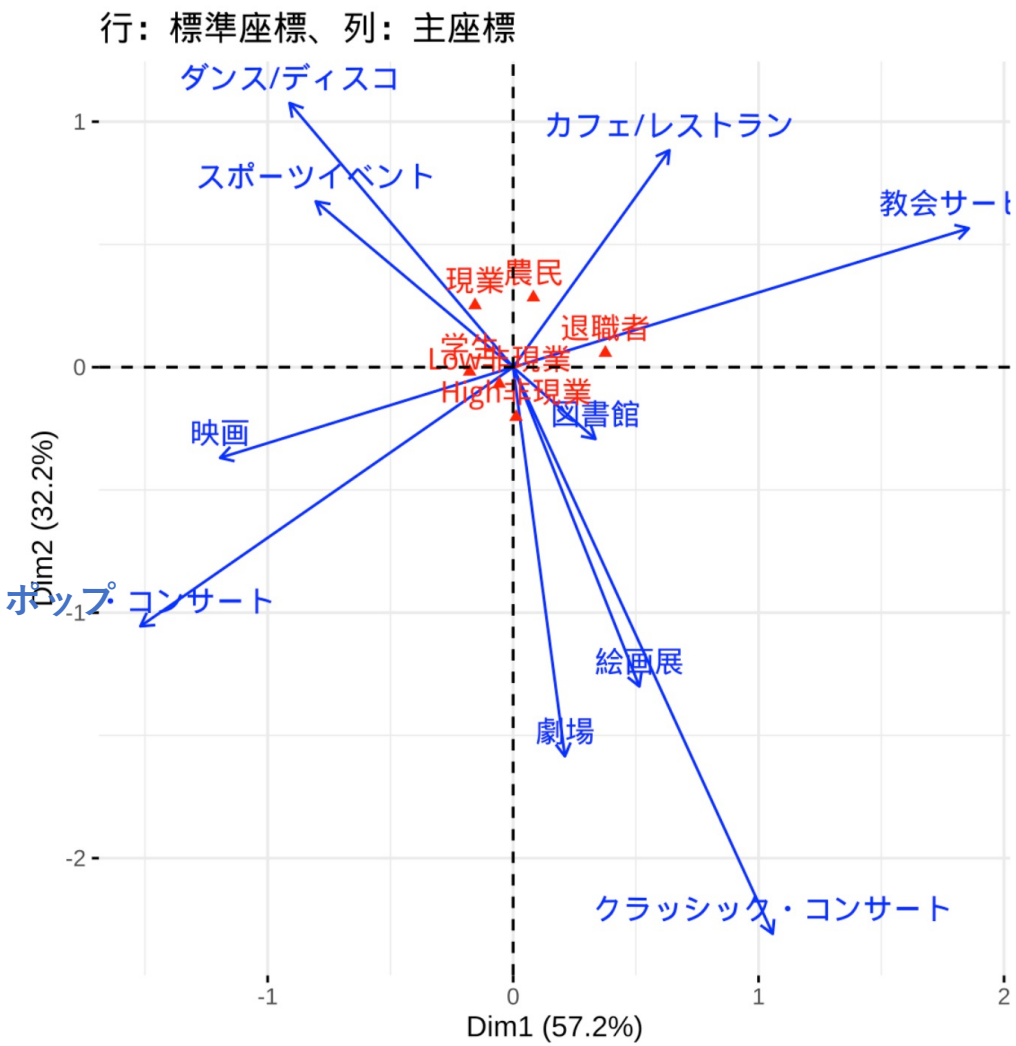
マップの見方（２）

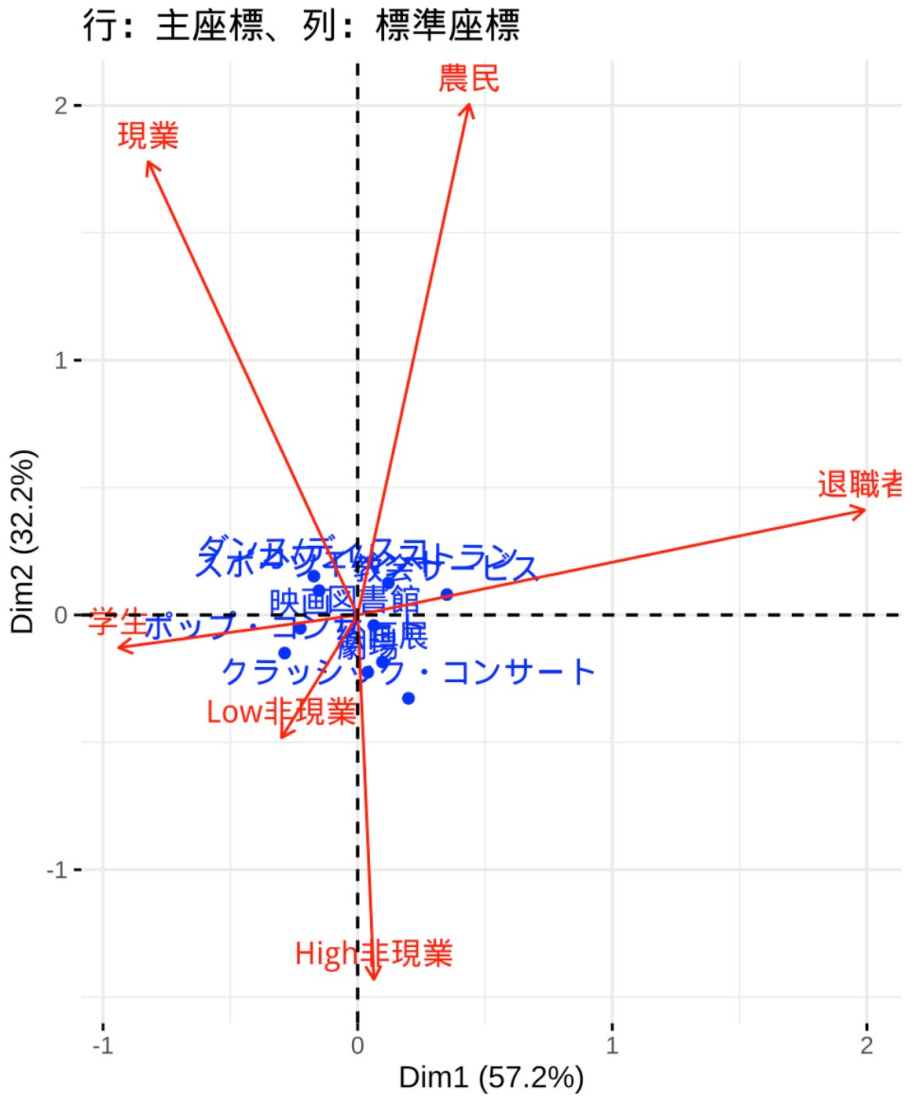
- 行変数内のカテゴリ間、列変数内のカテゴリ間、は定義されている。
- しかし、異なる変数のカテゴリ間の距離は定義されていない。

- ここが対応分析を理解する際の**ややこしい**ところ！

- 対策
  - 一方の変数を標準座標にして入れ物空間をつくり、そこにもう一方の変数カテゴリを射影する。非対称マップ。
  - 対称マップでは、このイメージをもって、位置ではなく、方向で考える。

非対称マップ


行：標準座標、列：主座標
ダンス/ディスコ
カフェ/レストラン
スポーツイベント
教会サービ
現業
農民
退職者
学生
Low非現業
High非現業
図書館
映画
ポップ・コンサート
絵画展
劇場
クラッシック・コンサート
Dim2 (32.2%)
Dim1 (57.2%)
1
0
-1
-2
-1
0
1
2

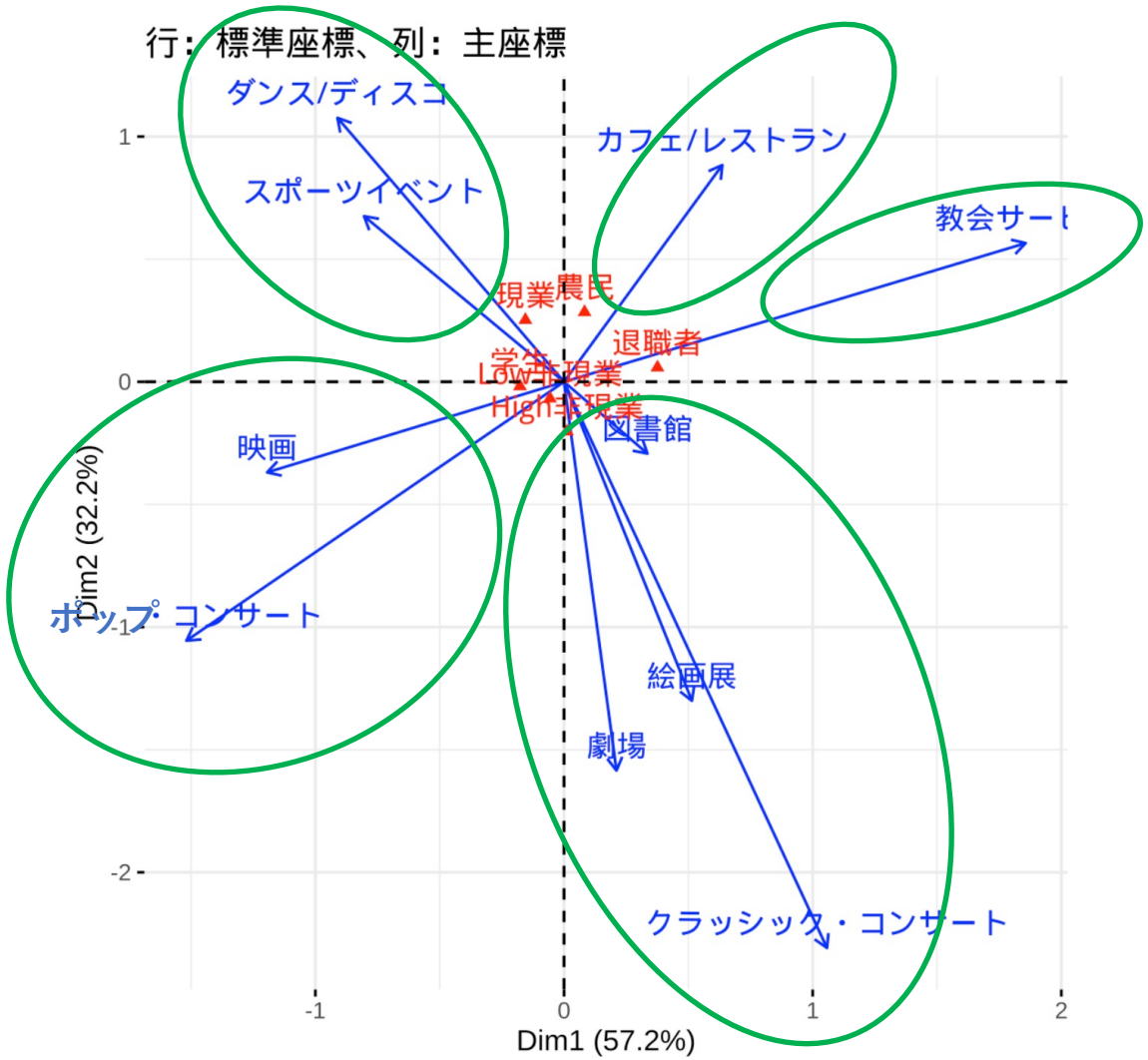

行：主座標、列：標準座標
農民
現業
退職者
学生
Low非現業
High非現業
クラッシック・コンサート
Dim2 (32.2%)
Dim1 (57.2%)
2
1
0
-1
-1
0
1
2

座標を解釈していく

- 第一軸 年齢
  - プラス方向：「退職者」
  - マイナス方向：「学生」
- 第二軸 職種での身体モード
  - プラス方向：身体性労働
  - マイナス方向：非身体性労働

- 他の解釈も可能かもしれません。
- 変数カテゴリの関係は、寄与率を表示して軸生成に寄与しているカテゴリに注目する。
  - 寄与率のグラフ化が有効

近いカテゴリを確認する



事例２ MCA(多重対応分析)

- SSM2005（社会移動と社会階層に関する全国調査2005）の留置A票の問16と、解答者の性別、年齢についてのMCA
  - SSM2005は、SSJDAのリモート集計で分析可能です。

問 16 男女の役割について、ア）～ウ）の意見があります。あなたはどう思いますか。それぞれについてあなたの考えにもっとも近い番号をひとつ選び、○をつけてください。

	そう思う	どちらかといえばそう思う	どちらかといえばそう思わない	そう思わない	わからない	
ア） 男性は外で働き、女性は家庭を守るべきである	1	2	3	4	9	⑪
イ） 男の子と女の子は違った育て方をすべきである	1	2	3	4	9	⑫
ウ） 家事や育児には、男性よりも女性がむいている	1	2	3	4	9	⑬

回答は、**1～4**でコーディングされていますが、カテゴリとして分析することを明示するために、**A～D、DKNA**でrecodeしてあります。

データのフォーマット

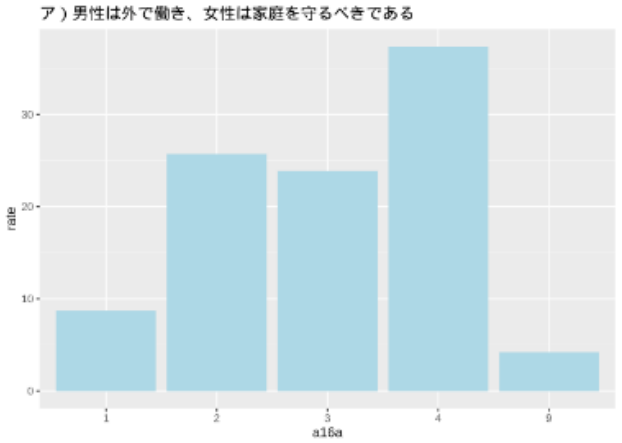
```
## # A tibble: 2,827 × 6
##      Age Age10 Sex    Q16a  Q16b  Q16c
##    <dbl> <fct> <fct>  <fct> <fct> <fct>
##  1    33 30    female D     C     B
##  2    55 50    male   B     C     C
##  3    47 40    female D     A     B
##  4    57 50    male   B     B     A
##  5    40 40    male   D     C     B
##  6    67 60    male   C     B     C
##  7    29 20    female D     D     D
##  8    37 30    female D     D     D
##  9    64 60    male   C     C     C
## 10    47 40    male   C     C     C
## # ℹ 2,817 more rows
```

```
summary(.d)
```

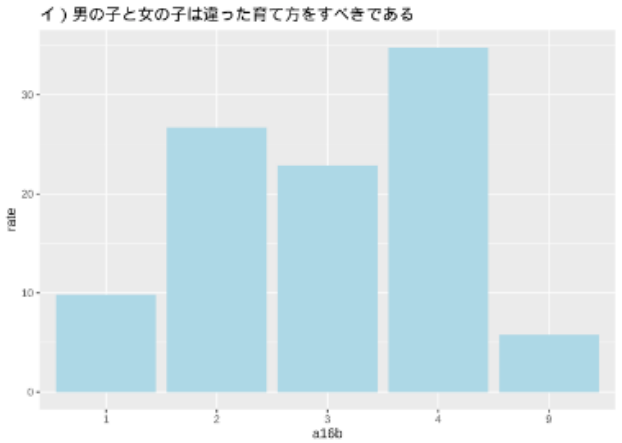
```
##       Age         Age10        Sex          Q16a        Q16b       Q16c
##  Min.   :20.00   20:325   female:1484   A   : 246   A   :277   A   : 572
##  1st Qu.:37.00   30:516   male  :1343   B   : 727   B   :754   B   :1141
##  Median :50.00   40:562                 C   : 677   C   :647   C   : 418
##  Mean   :48.17   50:684                 D   :1057   D   :985   D   : 545
##  3rd Qu.:60.00   60:726                 DKNA: 120   DKNA:164   DKNA: 151
##  Max.   :70.00   70: 14
```

変数ごとの単純集計（１）

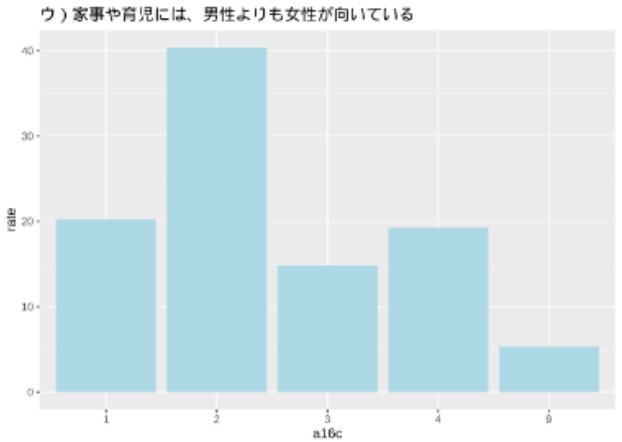
a16a	n	rate
1	246	8.7
2	727	25.7
3	677	23.9
4	1057	37.4
9	120	4.2



a16b	n	rate
1	277	9.8
2	754	26.7
3	647	22.9
4	985	34.8
9	164	5.8

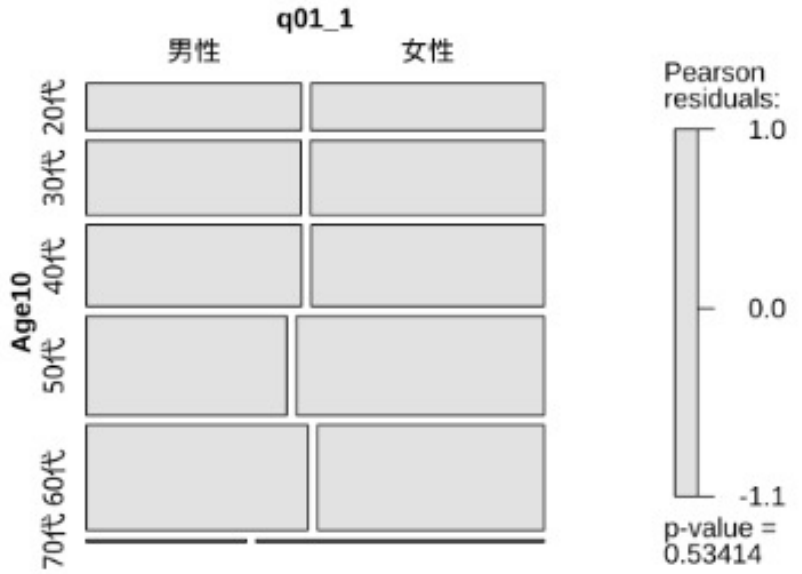


a16c	n	rate
1	572	20.2
2	1141	40.4
3	418	14.8
4	545	19.3
9	151	5.3



単純集計（２）年齢/年代と性別のクロス

	男性	女性
20代	156	169
30代	247	269
40代	270	292
50代	306	378
60代	359	367
70代	5	9

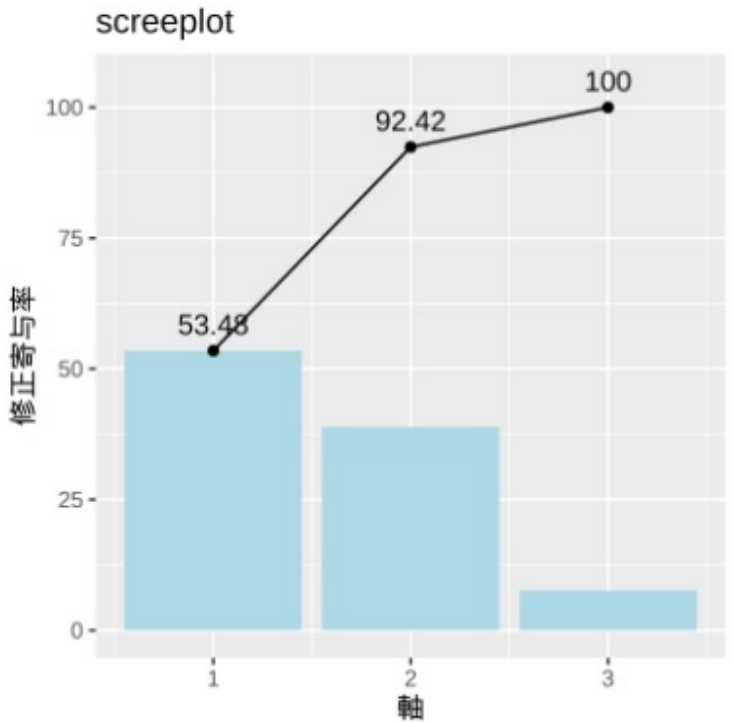


MCAを実行

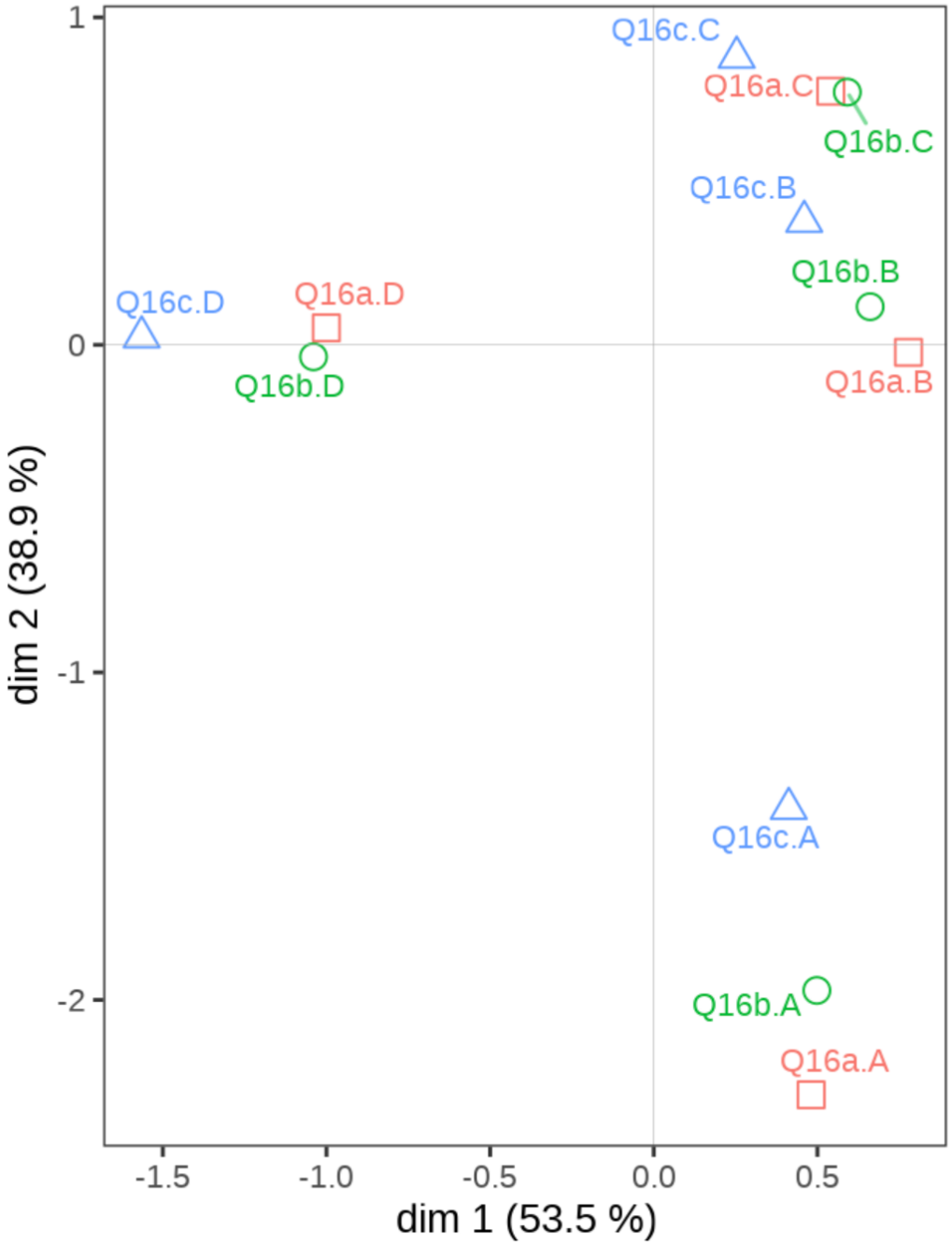
二軸で表全体の分散の92.42%を表現できる。

1－2次元分析でいく。

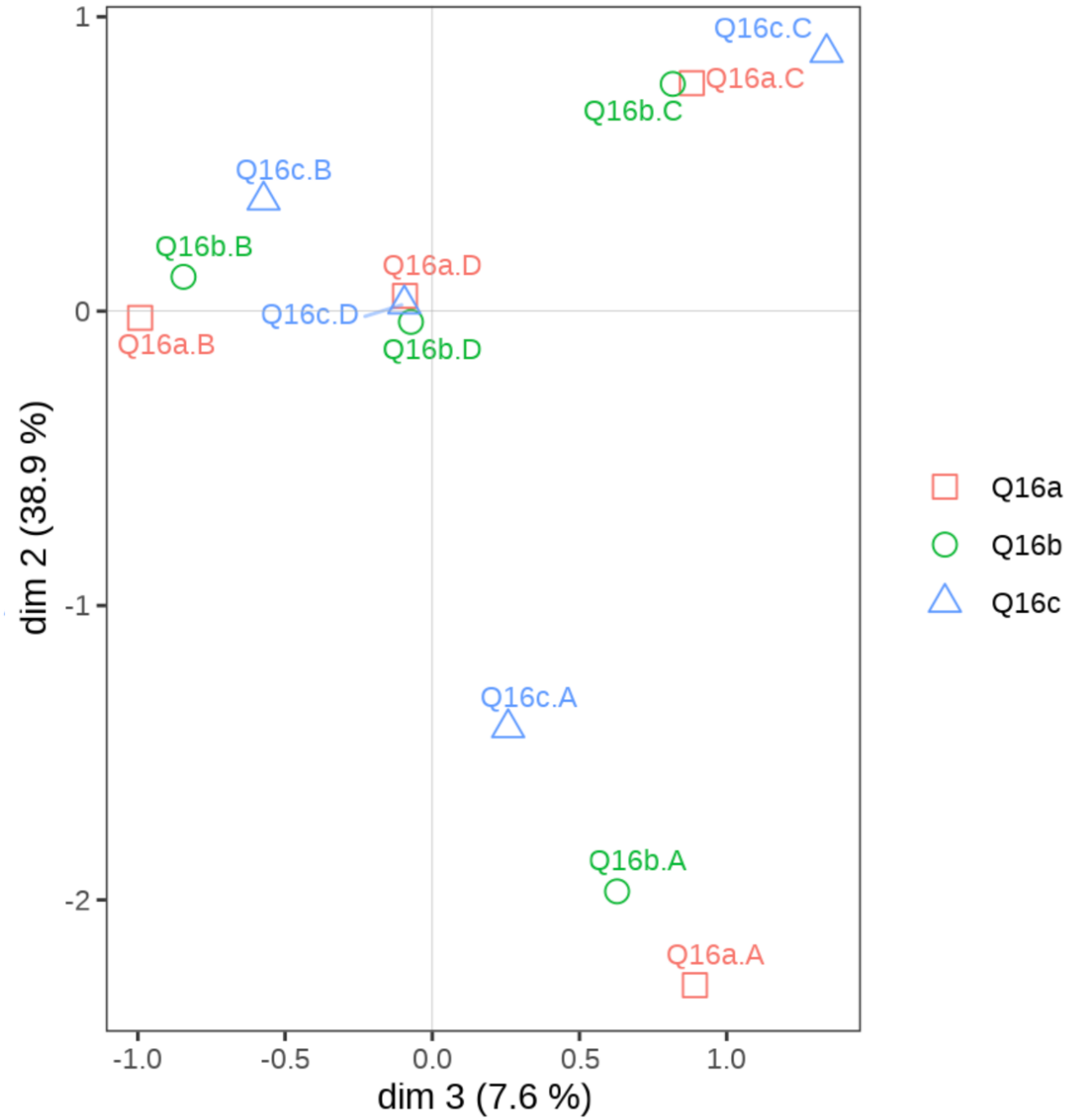
軸	修正寄与率	累積修正寄与率
1	53.48	53.48
2	38.94	92.42
3	7.58	100.00



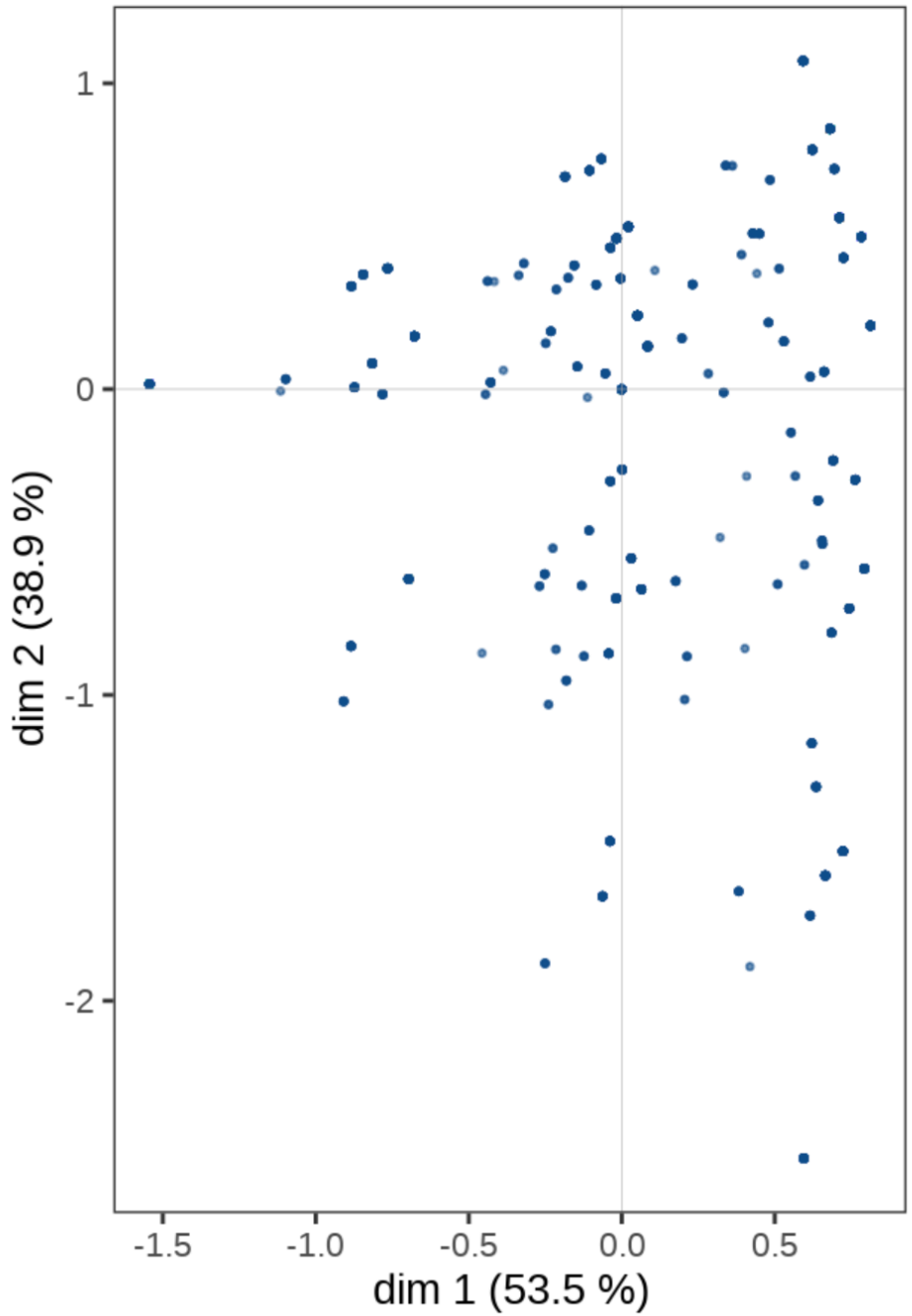
変数雲1−2軸



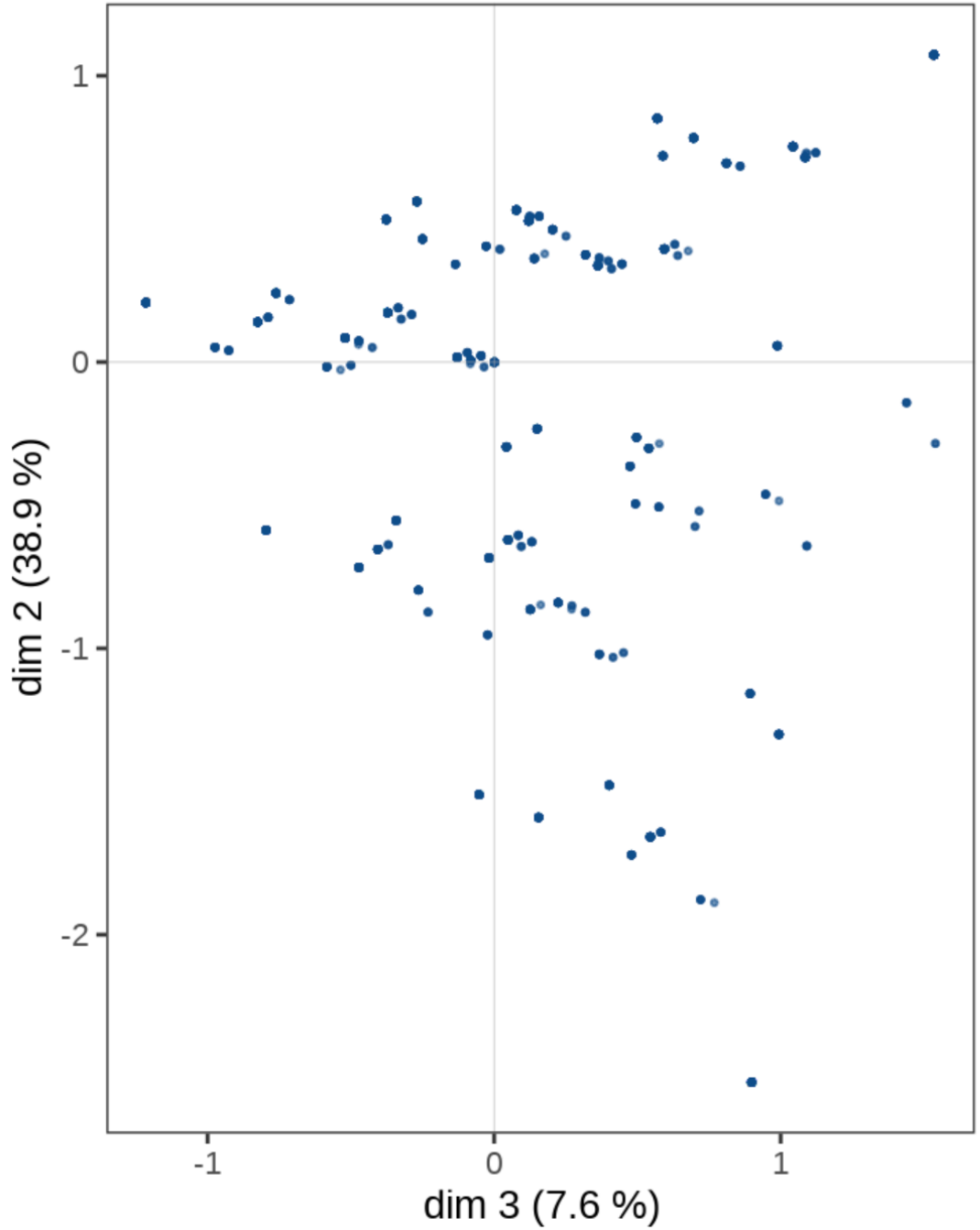
変数雲3−2軸



個体雲1-2軸



個体雲3−2軸



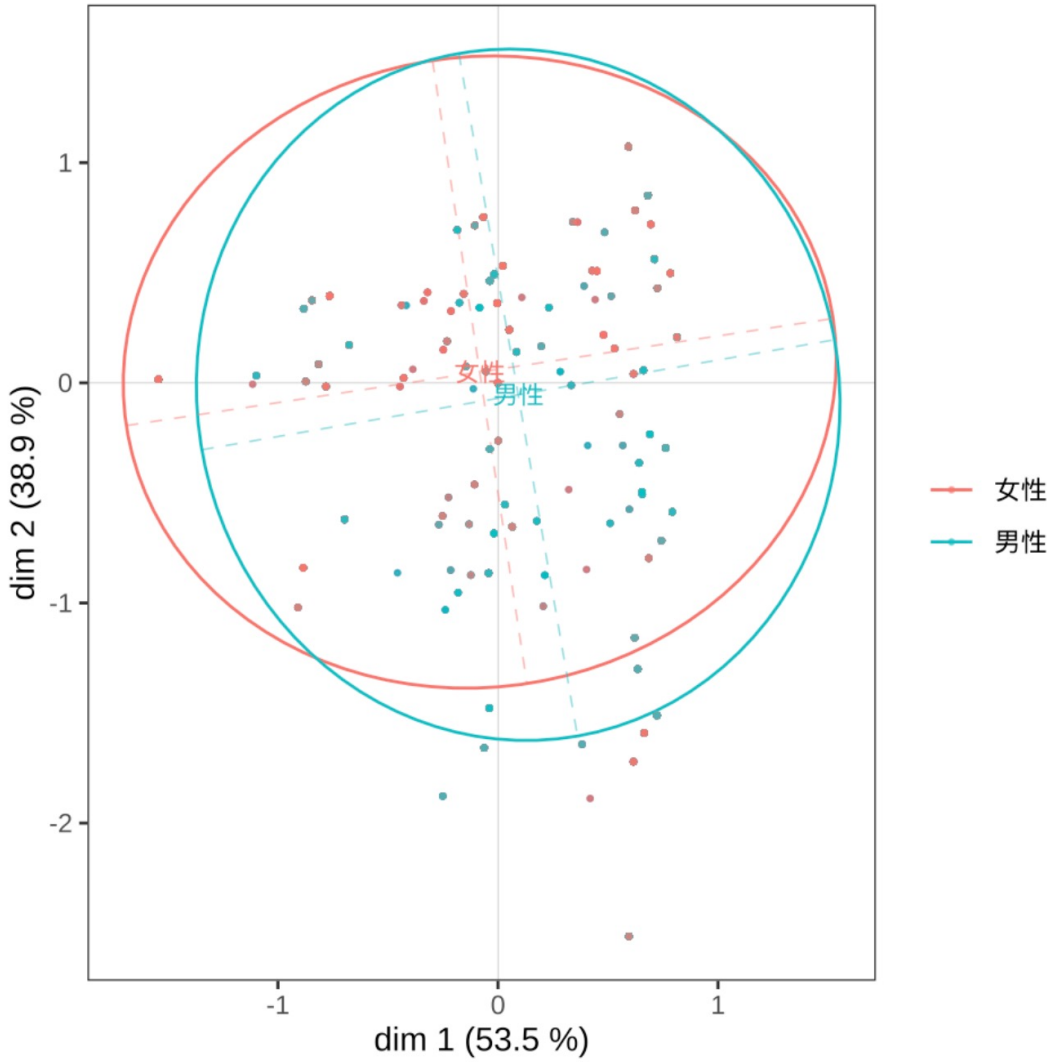
この展開への解釈

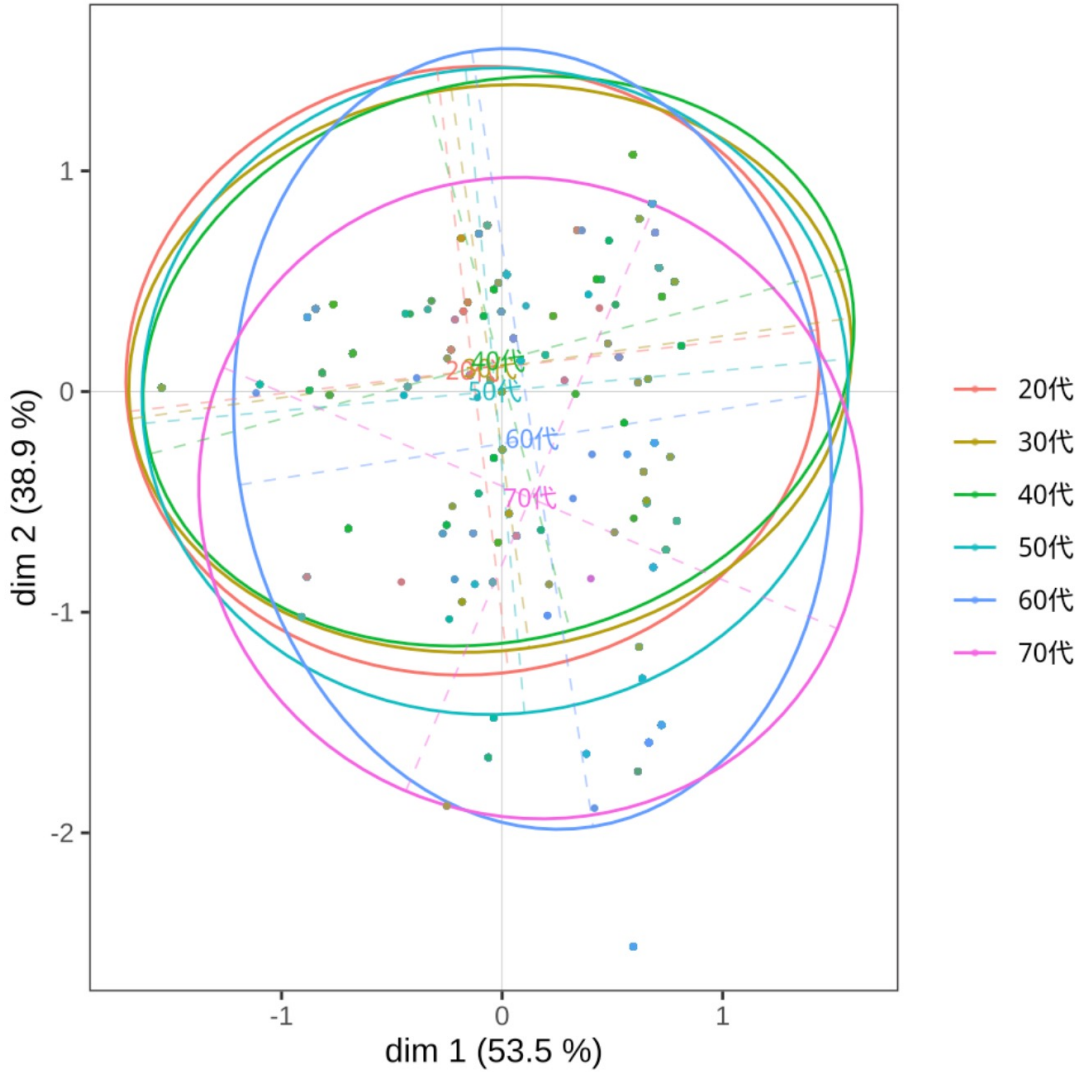
- まず、変数雲に注目して、生成された「軸」（これが新たな変数に相当します）を命名します。
- なにかしら自動的に（文化資本＋/経済資本ー）というような軸がでてくるわけではなく、分析者の責任で名付けます。
- この変数雲をみると
  - 第１軸
    - ーリベラル　＋どちらかと..と、保守が位置している。
  - 第２軸
    - ー保守。では縦方向の「差異」はなんだろうか。
- いずれにしても、A（保守）とD（リベラル）の間のB、Cは、近くにあるが、リニアではない。

個体雲に、追加変数を射影して空間分析

- 空間を生成（座標軸を生成）する変数とは別に、周辺度数をゼロにした変数を空間構造には影響をあたえずに、plotすることができる。（サプリメンタリ変数。追加変数）
- これを用いると、生成された空間を目的変数にみたてて、追加変数によって構造を分析することが可能になる。

個体雲に、性別、年代の分布を表示

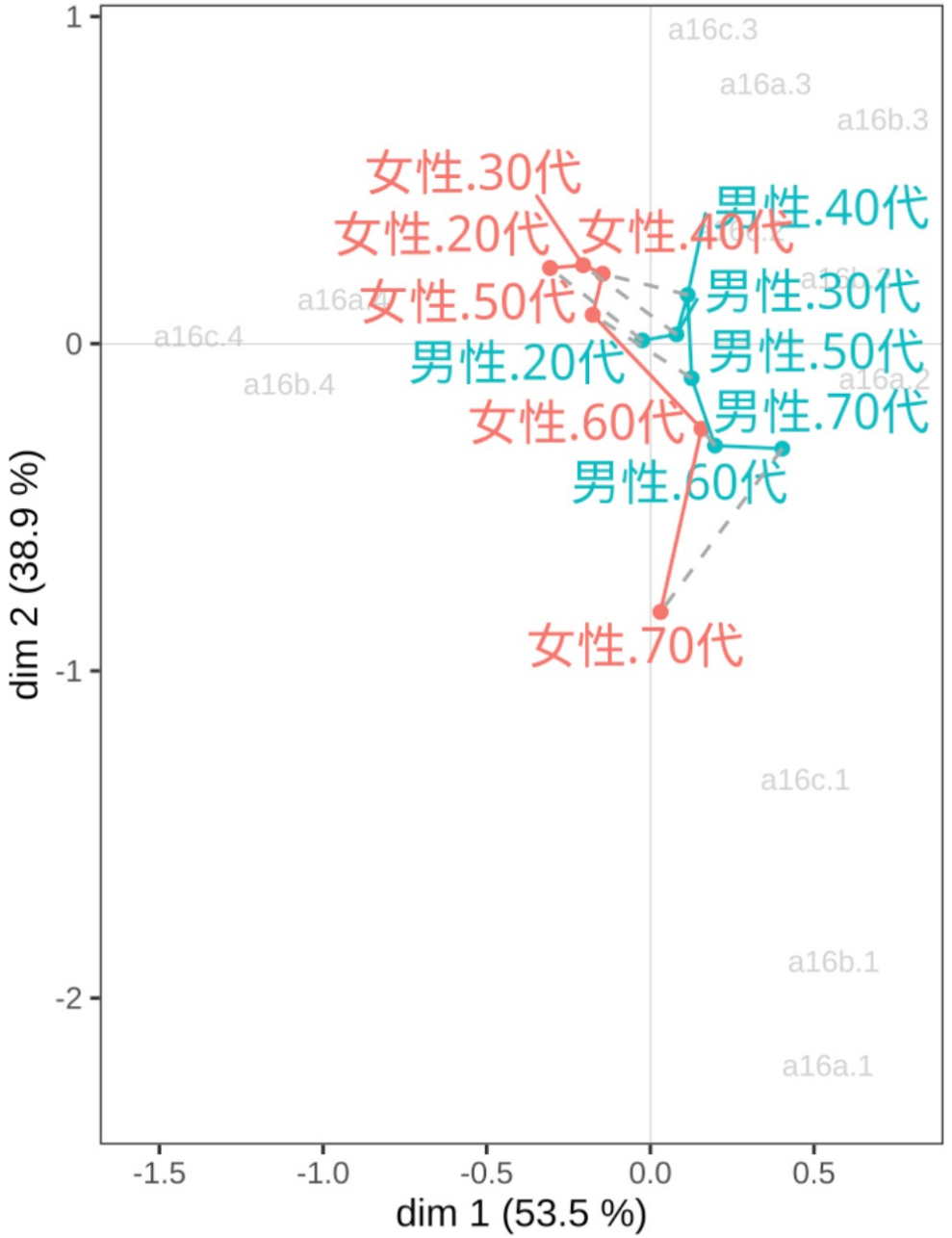




性別・年代の合成変数をつくり、交互作用を確認

性別（若年）は、第１軸の左右（マイナス側とプラス側）に分離しているが、年代が高いと女性も右側に入っている。

第２軸は、年代の若年ー高齢に対応。ただし男性・20代は別。
男女とも70代は、度数が非常にすくなかったので、60代に統合し、60代以上、とすべき。



CA/MCAの展開

- 原理的にはCAもMCAもシンプル。
- プロファイル間の距離をカイ二乗距離で評価して、次元縮減する。
- また、追加変数というアイデアが可能なので、生成した空間の多次元解析が可能。
  - これらの分析手法は、幾何学的データ解析として整備されている。（LeRoux & Rounaet 2004,2010=2021）

受講者「満足度調査」での活用

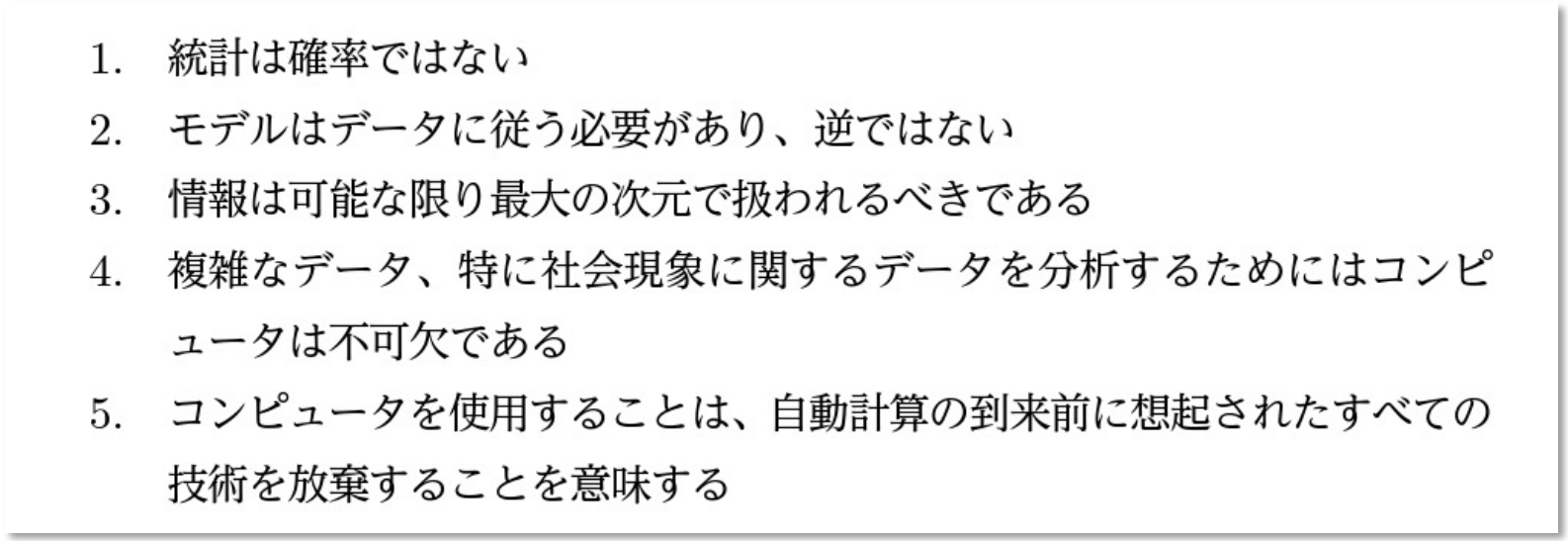
- 「ご祝儀回答」５、４、３、２、１で、ほとんどが、５と４。
- 回答選択肢に対する多重対応分析と自由記述部分回答に対するテキストマイニング、機械学習によるタグつけを行い。「ご祝儀回答」の中にうもれている、問題点の指摘、改善可能要素を抽出する手法を開発した。
  - NLP2023（言語処理学会2023）沖縄で発表。
  - **多重対応分析とアスペクトベース感情分析を組み合わせた受講者満足度調査データの分析手法の開発**
  - ○藤本一男, 大畑和也 (NICT)
  - https://www.anlp.jp/proceedings/annual_meeting/2023/pdf_dir/Q1-11.pdf

量的調査と質的調査（インタビュー）の連動（混合研究法）

- 個体が、平均値などに還元されすに、ポイントとして保存されている。そのために、マップ上で特徴的な位置（分布の隅っことか）に位置している個体のIDを取得して、インタビュー調査を実施できる。
- 例：Tベネット他（訳：磯他）『文化・階級・卓越化』青弓社

統計学の未来の姿をCAからみる

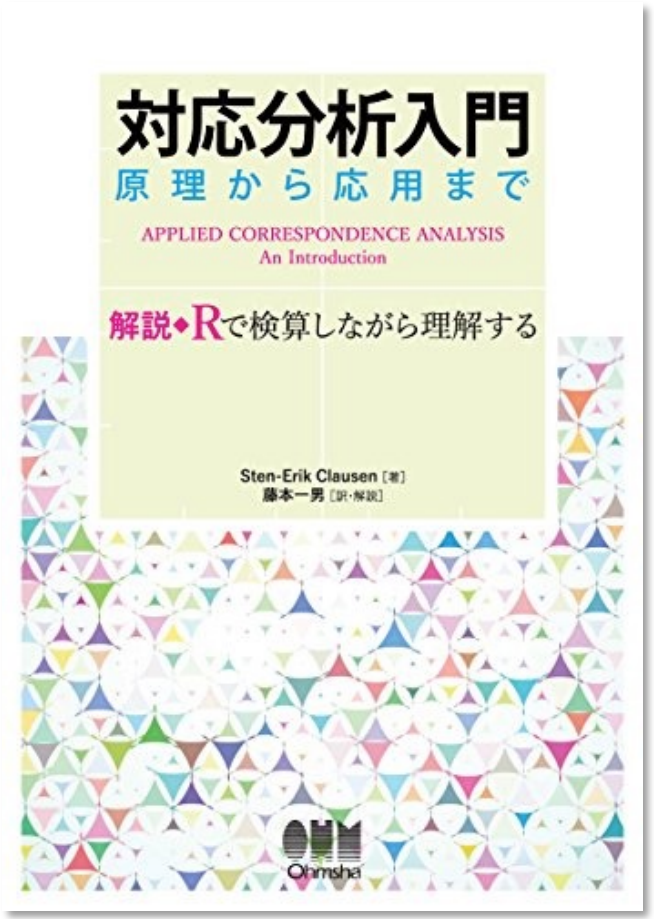
- CAをうみだしてきた、フランス学派、日本の林学派の統計学へのアプローチは、異色です。Greenacreの「日本語版への序」に面白いことが書いてあります。 https://419kfj.sakura.ne.jp/db/wp-content/uploads/2021/03/CAiP3%E6%97%A5%E6%9C%AC%E8%AA%9E%E7%89%88%E3%81%B8%E3%81%AE%E5%BA%8F.pdf



1. 統計は確率ではない
2. モデルはデータに従う必要があり、逆ではない
3. 情報は可能な限り最大の次元で扱われるべきである
4. 複雑なデータ、特に社会現象に関するデータを分析するためにはコンピュータは不可欠である
5. コンピュータを使用することは、自動計算の到来前に想起されたすべての技術を放棄することを意味する

参考文献

シンプルCAのみですが、この手法の概要がわかります。



MCAがタイトルですが、内容はGDA（幾何学的データ解析）の実践的解説書

CA/MCAに関する理論的解説。
応用を考える際に必要な理論解説はこちらで。

関連セミナー（2023/09/06）

- 東大社会科学研究所附属社会調査：データアーカイブ研究センター：CSRDAの「計量分析セミナー」2023のプログラムが公開されてます。
  - https://csrda.iss.u-tokyo.ac.jp/quantitative/seminar/
- 「対応分析」で講師をやります。
- シラバスは、ここにリンクされています。
  - https://csrda.iss.u-tokyo.ac.jp/9_6_2023summer.pdf

ご清聴、ありがとうございます！

本日以降でも、ご質問など、あればメールなどいただければご返信させていただきます。kazuo.fujimoto2007@gmail.com

Web: https://419kfj.sakura.ne.jp/db/