『津田塾大学紀要』第 56 号（2024 年）43-64 頁

帰納的データ解析（IDA）から見る「統計的検定」へのもう一つのアプローチ

藤　本　一　男

概要

MCA（多重対応分析）を中心にしたGDA（幾何学的データ解析）において構造化データ解析（SDA）までは、記述統計学に属する分析である。伝統的統計学では、統計学本来の目的は記述段階の次にくる推測段階であり、推測・検定が中心的に重要なものとして位置付けられている。しかし、このGDAでは、検定、推測は語られるものの、あくまでも、入手したデータに内在している意味をデータに語らせる、というアプローチをとる。

こうした視点を実現するものとして、SDAの次の段階のIDA（帰納的データ解析）では、典型性検定、同質性検定が実行されるが、その計算過程に現れたデータ分析観から統計的検定のもう一つのアプローチを明らかにした。

Keyword: 多重対応分析, 幾何学的データ解析, 帰納的データ解析, 統計的検定, 典型性検定，同質性検定，並べ替え検定，ブートストラッピング

1. 問題の所在

• MCA（多重対応分析）は推測、検定を扱えない？

MCA2010=2021:15 に「よくある質問」として「幾何学的データ解析でも、統計的推測をおこなえるか？」という問いかけが紹介されている。それへの回答としては、機能的に可能である、という回答に加え、統計的推測において重要なことはなにか、どのように推測を行うべきか、が重要であると、検定の前提にかかわる観点からの回答が述べられている。

• IDA（帰納的データ解析）自体が、知られていない

幾何学的データ解析（GDA）において、検定手法として「典型性検定

（typicality test）」「同質性検定」（homogeneity test）が提案されている（Le Roux & Rouanet 2004,2010=2021）。検定の問題状況としては、それぞれ、「1 群の平均値の差の検定」「2 群の平均値の差の検定」に対応しているのだが、典型性検定、同質性検定、と伝統的統計学の用語とは異なる名称が使用されている。それはなぜだろうか。そもそも、対応分析 / 多重対応分析が認知されていない上に、この手法が知られていないのが現状である。

• 帰納的データ解析（IDA）はどのような手法か。

そこで、本稿では、MCA2010=2021 の第 5 章の説明をもとに、組合せ論枠組みでの帰納的分析（IDA、記述統計からの推定・検定）を解説し、R を用いて伝統的検定手法も含めて、統計検定量、*p*- 値の比較を行うことでこの手法の特徴を明らかにし、それを通して、IDA という手法の背後にあるデータ分析観を明らかにしていきたい。GDA という手法については、藤本 2020、2022、2023 を参照。

なお、本稿での MCA の実行に用いたのは、Nicoras Robette, GDAtools2.0 である [Robette 2023]。

2. 対応分析（CA）による数量化のプロセス

検定対象は、MCA の結果を分析して明らかになった「差異」である。そのために、本題に入るまえに MCA での処理を簡単に整理しておく。（なお、CA/MCA の原理を把握している方は、3 帰納的データ解析（IDA）の考え方に進んでいただいてかまわない。）

CA /MCA は、二元表を分析対象とする。そこでは、行も列も、全体に対する割合を要素としたプロファイル･ベクトルとして表現され、CA /MCA は、そのプロファイルの関係をカイ二乗距離として計量化する。このプロファイル・ベクトルが、行列空間、列空間を生成し、そこでは、当初の n 次元空間が、少数次元に縮減される。この次元縮減によって生成される空間の座標軸は、変数カテゴリが統合変換されたものである。

こうして、生成された空間に対して、空間生成に寄与しなかった変数を「追加変数」として射影することが可能で（遷移公式 Clausen1998=2015:22, Greenacre 2017=2020:108）、これを用いて、まず、変数空間の構造分析が行われる。

次に、個体空間に対してこの追加変数を射影し、追加変数を説明変数とし

て、個体空間の分析が可能になる（構造化データ解析 SDA）。

この段階（記述統計）で明らかになった「差異」の有意性の分析が帰納的データ解析（IDA）として行われる。

この段階は帰納的データ解析（IDA）とよばれ、そこでは、組合せ論的枠組み（Combinatorial frame work）での分析が行われる。

そこでの検定は 1）典型性検定（Typicality test）と 2）同質性検定（Homogeneity test）と呼ばれる。伝統的検定技法と対比させるならば、1）は、「1 群の平均値の検定」、2）は「2 群の平均値の差の検定」である。

従来手法で同じ問題領域を扱うには、正規性の確認、等分散性の確認が必要になるが、この 2 つの検定では、後述のようにこの仮定は前提とされない。

そこで、IDA の特徴を明らかにするために、伝統的な検定手法を含めて比較する。

表 1　分析対象のデータセット（一部）「嗜好データ」

ID <int>	Isup <fctr>	TV <fctr>	Film <fctr>	Art <fctr>	Eat <fctr>	Gender <fctr>	Age <fctr>	Income <fctr>
1	Active	TV-メロ…	映画-アクション	芸術-風景画	外食先-ステーキハ…	女性	55-64	£20-29
2	Active	TV-メロ…	映画-ホラー	芸術-静物画	外食先-インド料理店	女性	45-54	<£9
3	Active	TV-自然	映画-アクション	芸術-風景画	外食先-パブ	女性	55-64	<£9
4	Active	TV-メロ…	映画-時代劇	芸術-肖像画	外食先-イタリア料…	女性	65+	£10-19
5	Active	TV-コメ…	映画-ホラー	芸術-静物画	外食先-インド料理店	女性	35-44	£10-19
6	Active	TV-コメ…	映画-ホラー	芸術-印象派	外食先-インド料理店	女性	18-24	<£9
7	Active	TV-ニュ…	映画-アクション	芸術-風景画	外食先-インド料理店	女性	25-34	£10-19
8	Active	TV-ニュ…	映画-ドキュメ…	芸術-パフォーマン…	外食先-フィッシュ…	男性	65+	£10-19
9	Active	TV-メロ…	映画-時代劇	芸術-風景画	外食先-ステーキハ…	女性	65+	<£9
10	Active	TV-ニュ…	映画-アクション	芸術-風景画	外食先-フィッシュ…	女性	65+	£10-19

1–10 of 1,253 rows　Previous 1 2 3 4 5 6 … 100 Next

2.1. 構造化データ解析（SDA）による記述統計学的分析

伝統的な統計学の教科書では、統計学は「記述統計学」と「推測統計学」にわかられ、前者は、手元にある収集されたデータについて分析すること、であり、後者は、その手元のデータの分析をとおして、手に入れてないデータ（たとえば母集団）のデータを推測すること、とされる。この両者をつなぐものとして「確率論」が位置している。

GDA における SDA（構造化データ解析）までは、この「記述」段階に相当する。この記述段階で明らかになった注目すべき「差異」に対して、それが、偶然生じているのか、それとも生じるべくして生じているのかを明らかにするために、検定が行われる。

データに対して、MCA を実行するにあたって、個体空間、変数空間を生成するアクティブ変数（Active variable）と、それを「目的変数」として説明する「追加変数」（Supplementary variable）への分割が行われる（構造設計）。本稿で事例として扱う「嗜好データ」[1)]では、TV（TV）、映画（Film）、芸術（Art）、外食（Eat）の4変数がアクティブ変数として空間生成をにない、性別（Gender）、年齢（Age）、収入（Income）の3変数が追加変数として、生成された空間の構造を解釈するために用いられる。このように構造設計された変数を用いて MCA を実行すると、以下のような空間が二つ生成される（ここでは 1 − 2 軸平面のみ提示）。

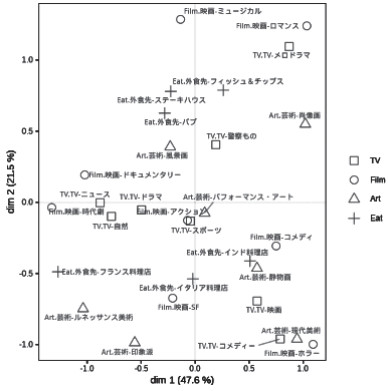

図1　嗜好データ MCA、変数空間

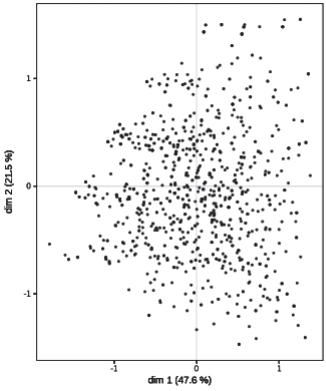

図2　嗜好データ MCA、個体空間

変数空間においても個体空間においても、似たものは近くに、異なったものは遠くに配置されている。原点（O）は、全体のデータの平均点、つまり帰無仮説に対応する位置にある。それゆえに、原点、また水平軸、垂直軸を境目として反対側に位置しているものは、データが示す平均に対して対照的に異なった性格を有していると解釈できる。

ところで、生成された空間の各軸は何を意味しているのだろうか。各軸は、元のデータ表の変数カテゴリ（この例は 29 カテゴリ）が張っていた多次元空間を次元縮減によって、低次元で近似したものである。この軸は、MCA によって変換（数量化）された新たな「変数」なのである。

それゆえ、この二つの空間を前に、まずやるべきなのは、軸に名前をつけることである。まず変数空間で命名し、個体空間でもそれをもとにデータ構造の分析を進める。

2.2. 各変数カテゴリの寄与率から、軸への意味付与を行う

各変数カテゴリごとに、軸の生成（新変数の生成）に寄与している。

軸の命名は、変数空間に注目して行う。そこでの各変数カテゴリの軸への寄与率を見ることで、軸のマイナス方向、プラス方向にどのカテゴリがどのくらい寄与しているのかを確認することができる。プラス / マイナスの方向は、座標値で確認でき、カテゴリ数から得た平均値より大きな寄与率に注目する（Le Roux, Rouanet 2010:2021:70）。

この値は、**GDAtools::dimcontriv**（）で得ることができる（表2）。また、変数空間マップを表示する際に、**ggcloud_variables**（）で、**point=best**、**besth**、**bestv** を選んで、注目すべきカテゴリだけを表示することができる（図3）。さらに、GDAtoolsV2.0 から追加された機能に、**ggaxis_variable** という機能があり（図4）、軸ごとに寄与率に対応した文字サイズで表示してくれる。軸の命名の第一段階では、これらのツールを活用する。

表2　各変数カテゴリの軸への寄与

	Variable	Category	Weight	Quality of representation (left)	Contribution (right)	Total contribution	Cumulated contribution	Contribution of deviation	Proportion to variable
9	Film	映画-時代劇	140	0.230	12.69	34.2	34.2	33.16	95.81
5		映画-コメディ	235	0.135	6.79				
8		映画-ロマンス	101	0.097	5.55				
6		映画-ドキュメンタリー	100	0.094	5.37				
7		映画-ホラー	62	0.064	3.8				
11	TV	TV-ニュース	220	0.172	8.78	26.9	61.1	26.81	87.2
12		TV-メロドラマ	215	0.163	8.37				
13		TV-自然	159	0.090	4.91				
10		TV-コメディー	152	0.089	4.85				
4	Eat	外食先-フランス料理店	99	0.143	8.21	13.55	74.65	12.92	84.2
3		外食先-インド料理店	402	0.128	5.34				
2	Art	芸術-肖像画	117	0.111	6.26	11.28	85.93	11.26	58.33
1		芸術-現代美術	110	0.088	5.03				

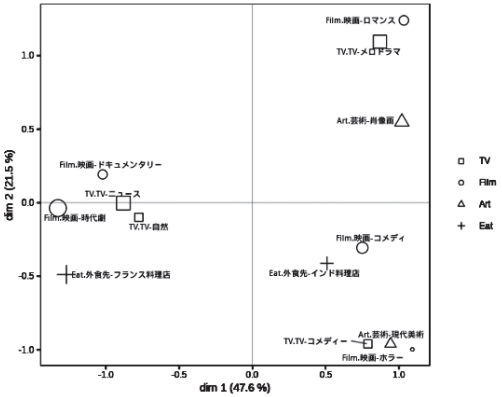

図3　軸1に大きく寄与している変数カテゴリ（平面）

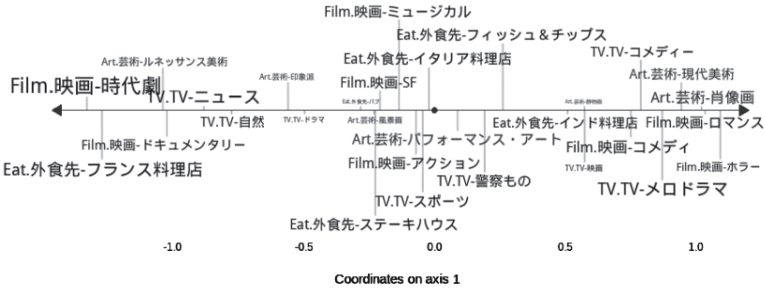

図４　軸１に寄与している変数カテゴリ（1 軸上カテゴリ）

これらのツールを用いて、軸ごとに命名したものが以下のものである。この軸の解釈、命名は機械的には行えず、分析者の分析対象に対する専門的な知見にもとづいた解釈となる。ここにあげたのは、MCA2010=2021 にある例であるが（Le Roux&Rouanet2010=2021:72,73,74）、まったく別の命名が排除されるわけではない。

第 1 軸：（＋）事実に即したもの / 伝統的なもの ⇔（－）架空のもの / 現代的なもの。

第 2 軸：（＋）大衆的 ⇔（－）洗練されたもの

第 3 軸：（＋）硬いもの ⇔（－）柔らかいもの

なお、軸の命名は、こうした寄与率を基準におこなわれる。そのため、論文などに報告する際は、空間マップ、と注目すべき変数カテゴリの寄与率を表として掲載する、という「慣例的ルール」がある。

2.3. 性別、年齢、収入を追加変数とした構造分析

MCA のリザルトの個体変数の座標部分のデータに、追加変数（Gender: 性別、Age: 年齢、Income: 収入）を接続すると表 3 のようなデータ構成となる。ここで dim.1-dim.3 が個体の座標値。Gender、Age、Income が追加変数である。この追加変数のカテゴリを用いて、個体空間の部分空間への分割が可能になる（例えば、男性個体の部分空間。女性個体の部分空間）。

表3　MCA リザルト（座標）と追加変数の結合

	dim.1 <dbl>	dim.2 <dbl>	dim.3 <dbl>	Gender <fctr>	Age <fctr>	Income <fctr>
1	-0.1353093622	-0.901984475	0.4323046887	女性	55-64	£20-29
2	-1.2024365556	0.328563329	-0.2638075182	女性	45-54	<£9
3	0.5370106529	-0.333733358	0.5648763197	女性	55-64	<£9
4	-0.2136401677	-0.451225657	-1.1141370254	女性	65+	£10-19
5	-1.1699785032	1.195527931	-0.0659147224	女性	35-44	£10-19
6	-0.7225735345	1.416187118	-0.3761584224	女性	18-24	<£9
7	0.2664249744	0.064106352	0.4380746986	女性	25-34	£10-19
8	0.6140031079	-0.380673096	0.2981257886	男性	65+	£10-19
9	0.3615201602	-0.940031539	-0.3982101637	女性	65+	<£9
10	0.3641543297	-0.442372721	0.5231054790	女性	65+	£10-19

1-10 of 1,215 rows　Previous 1 2 3 4 5 6 ... 100 Next

性別、年齢について、集中楕円（concentration ellippses）[2)] を生成すると、その中心点座標が全体の原点（全体の平均点）からずれていることが確認できる。

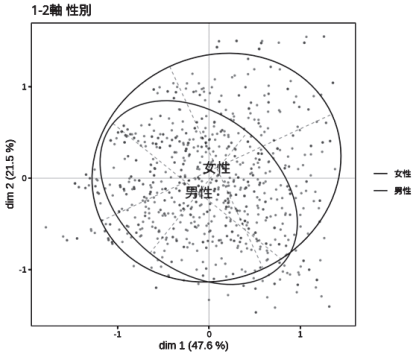

図5　追加変数の集中楕円（性別1-2次元）

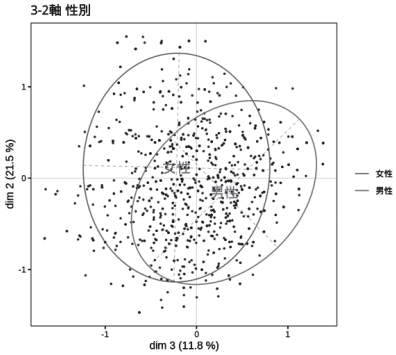

図6　追加変数の集中楕円（性別3-2次元）

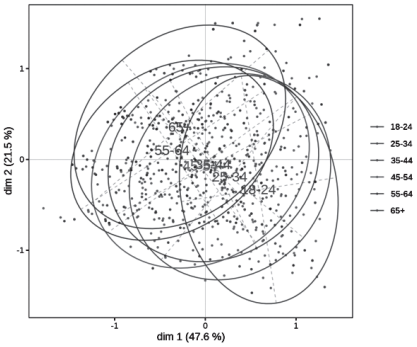

図7　追加変数の集中楕円（年齢1-2次元）

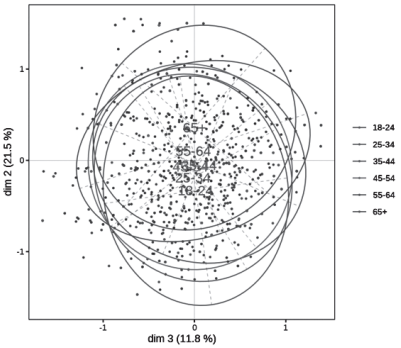

図8　追加変数の集中楕円（年齢3-2次元）

構造化データ解析（SDA）段階、つまり記述段階ではこの差異について、分解された分散とその比（η^2）を手掛かりに、分析、解釈を行う。平均値という座標値の差異だけでなく、その部分集合のひろがり / 散らばりも分析の差異の基準とする必要がある。

次の図は性別についての座標値（図9）と群内分散 ***Vwithin***（図10）の違いを表示したものである。これを見ると、座標値の男女の違いは、3軸において顕著であり、群内分散 ***Vwithin*** でみると、そのバラツキは1軸2軸で顕著であると解釈される。

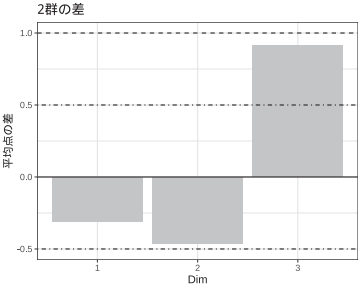

図9　座標値の原点との差

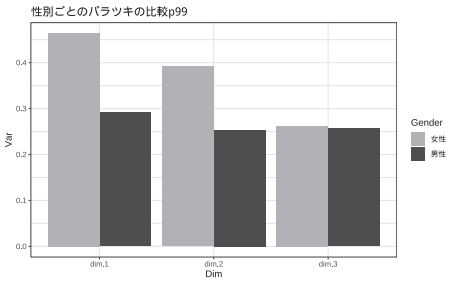

図10　分散値 ***Vwithin*** の比較

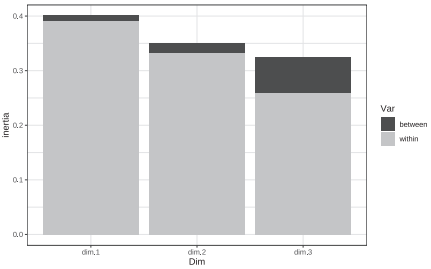

図 11　軸慣性（inertia）内の η^2 値の比較

表 4　座標値と η^2 との差異

	dim.1	dim.2	dim.3
男性	0.291503	0.252755	0.256650
女性	0.463911	0.391608	0.261315
within	0.391117	0.332981	0.259345
between	0.009239	0.018185	0.065664
total	0.400355	0.351166	0.325009
eta2	0.023076	0.051784	0.202037

$$\eta^2 = \frac{Vbetween}{Vtotal}$$

η^2 を見ることの重要性

平均値の差の大きさだけでなく、分散も評価する必要があることはいうまでもないが、ここで、相関比 η^2（群間分散（Vbetween）/ 全体分散（Vtatal））の比較をした図 11 を見て欲しい。このように、カテゴリのデータ構造での影響度合いを解釈するには、平均値の差、分散（集中楕円）、分散の比率（η^2）を総合的に評価し解釈する必要がある[2)]。

その結果は、以下のようにまとめられる（Le Roux & Rouanet 2010=2021:xx ）。

第 1 軸　年齢に関係している
第 2 軸　収入と年齢に関係している
第 3 軸　性別に関係している。

このように、軸の解釈は多面的に行われることになる。このようにして「差異」が明らかになった次の問題は、ここに記述的に確認された差異は有意なのか否かである。

3. 帰納的データ解析（IDA: Inductive data analysis）の考え方

Le Roux&Rouanet2010=2021 第 5 章は、組み合わせ論的枠組みで典型性検定（Typicality test）と同質性検定（Homogenity test）を説明する。ここでいう二つの検定は、一般的な統計学が設定する問題状況でいえば以下のものに対応している。

典型性検定（Typicality test）

注目している追加変数による部分集合の平均点がデータ全体の平均（これは原点）からずれている。このズレずれが、ずれるべくしてずれている

のか、偶然ずれているのか、によって、部分集合が全体雲の典型（typical）か否（atypical）かを判断する。

対応する問題状況は、「1群の平均点の比較対象μとの差」の検定である。

同質性検定（Homogenity test）

典型性検定は、注目した部分集合と平均点がゼロ（原点）である全体雲との関係を問題にしているが、同質性検定は、2つの部分集合に注目する。2つの部分集合の平均点にずれがある場合に、そのずれが有意かどうかを検定する。

問題状況としては、「2群の平均点の差」の検定である。

Le Roux らが強調するのは、この手法で検定する場合は、伝統的手法では必要とされる仮定、すなわち「分布の正規性」、2群の差の検定であれば、加えて「等分散性」が、必要とされないということである。それは原理的にはMCAのリザルトを参照母集団（reference population）としてリサンプリング（並べ替え）によって生成される標本分布は、中心極限定理での標本分布の展開そのものであり、正規性への漸近を強く主張できるからである。

組み合わせ論的枠組み（Combinatorial Framework）

IDAでは、SDAで得られた「差異」を組み合わせ論的枠組みにそって考えていくので、まず、その考え方を整理しておく。

なお、以下の主軸における典型性検定の説明においては、追加変数はAge（年齢）を、また、そのなかでも第一軸（dim.1）を対象としている。また、平面における典型性検定においては、dim1とdim2を用いるが、その場合の並べ替え分布の生成については、図A.2を参照。

検定する対象は、MCAのresultとして取得された変数空間、個体空間という二つの空間の座標に体現されている軸ごとの分散（慣性）と個体の各点の座標（主座標）である。MCAの実行によって各変数カテゴリの分布は、以下の図12のように得られた。この分布をもとにしてこのあとに述べる並べ替えリサンプリングによって、図15のような分布に変換される。このグラフに表現されているのは、年齢（Age）で区分された部分集合ごとの第一軸の座標値の分布であるが、加えて、検定対象としてこれらの分布の平均値がそれぞれ存在している。

検定の課題は、その平均値が、それぞれの分布からみて、全体の分布の平均値であるゼロから有意に離れているかどうかである。

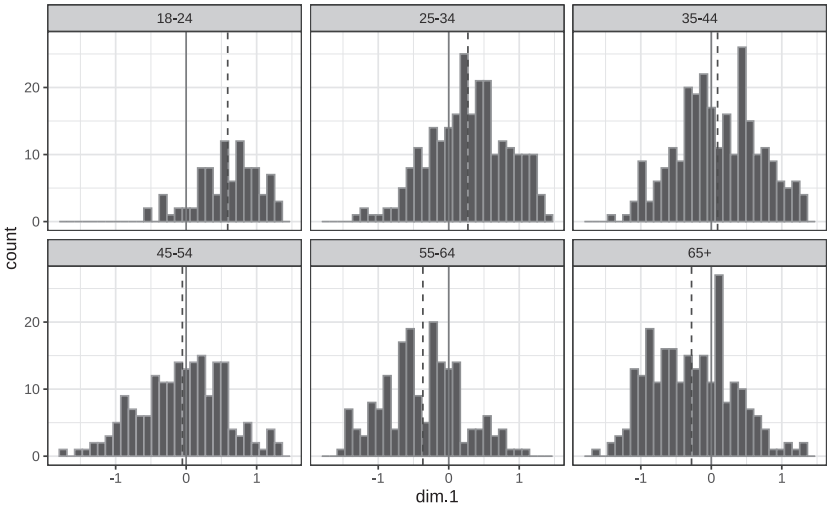

図 12　MCA 年齢部分集合の分布（第 1 軸）

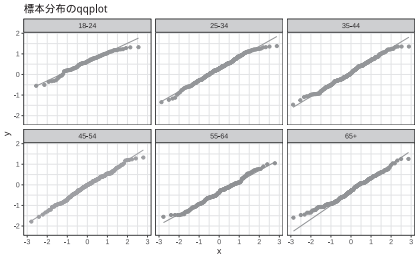

図 13　年齢 QQplot（正規性確認）第 1 軸

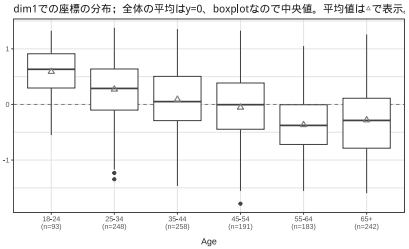

図 14　年齢の箱ヒゲ図（第 1 軸）

IDA はこの分布から非復元抽出リサンプリング（図 A.1）によって、次のグラフ（図 15）に見られるような分布へと変換する。検定統計量の値（座標値）は、グラフに波線で書き込まれている。

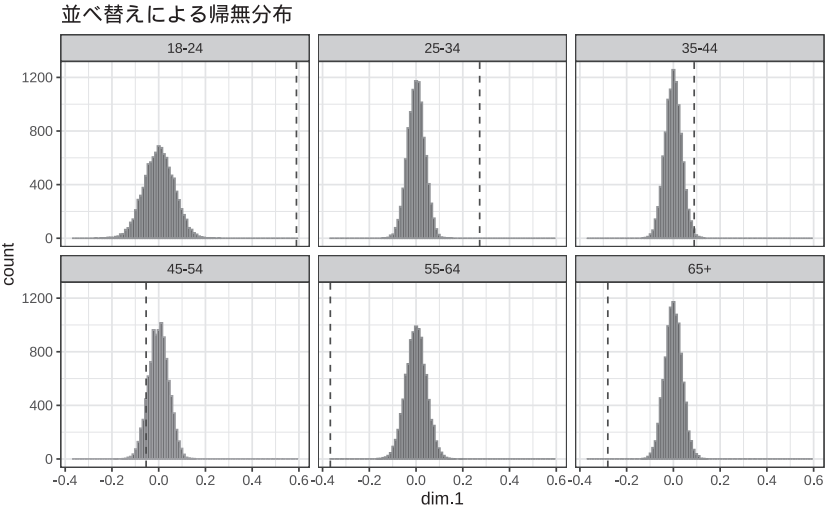

図 15　並べ替えによる年齢部分集合 1 軸の帰無分布

組み合わせ論枠組みにおける検定の考え方も、基本は伝統的な検定の考え方と同じである。検定統計量を帰無分布との関係で評価する。つまり、確認された分布（これは入手されたデータからランダムに抽出された標本の平均値の標本分布である）からみて、垂線の位置にある値は、たまたま実現された値（つまり差は偶然）なのか、それともしかるべき理由が存在して実現された値（つまり差は有意）なのか、を問題にする。

そもそも MCA を用いずにこれらの関係を分析するとしたら、度数分布表を作成しカイ 2 乗検定によって、年齢群によって分布に違いがあるかということしか見ることはできない。今展開している検討が可能なのは、MCA によって、元のデータの数量化が行われたからである。

そこで問題は、MCA 後に取得された図 12 の各分布とそれぞれの平均値との関係である。全体の分布の平均は、ゼロ（原点）であることがわかっているので、図 12 に対して、検定を行うとした場合は、各分布の平均点の座標値と原点ゼロの差が有意か否かという検討になる。

伝統的な t- 分布を用いて、図 12 の分布を分析することを考える。その平均値と原点との差について t- 検定を行うことになる。これを方法 0 としておく。

この方法を実行するにあたっては、検討すべきことがある。それは、検定対象の分布が正規分布をしているか否かである。これについては。シャピロ・

ウイルク検定を行い、またQQplotを描画し確認するか、ノンパラメトリック検定としてWelchの順位和検定を行うことになる。

この正規性の仮定は、明確なかたちで確認されることはない。そもそもシャピロ・ウイルク検定を始めとして正規性の検定は、帰無仮説が「正規分布である」というものである以上、もしp値が0.05より大きな値をとったとしても積極的な主張として「正規分布である」ということはできない。帰無仮説の性格からして、せいぜい「正規分布ではない、とはいえない」と主張できるだけである。

対して、組み合わせ論的な検定では、以下のようにアプローチすることでこの伝統的検定手法がもとめる前提的仮定の必要性を回避する。MCAの結果取得された部分集合ごとの分布を「参照母集団」（reference population）として、そこからのリサンプリングによって、並べ替え分布を生成する。それは中心極限定理によって、正規分布（もしくはχ^2分布）への漸近が保証されており、確率密度関数による「近似」が強く保証されている（図A.1）。

図12にみられる分布のデータに対して、並べ替え分布を計算し、それと平均値の位置を確認される。

この並べ替え分布（帰無分布）の取得には、多くの場合実現が困難な厳密版（方法1-0）と二つの近似方法（方法1-1、1-2）がある。

1-1）図12の分布のデータから非復元抽出によって、R個（本稿では10,000個）の標本データを生成し、それを帰無分布とする（モンテカルロ近似）。

1-2）1）の非復元抽出によって生成される分布は中心極限定理によって、正規分布近似が可能になるので、それを帰無分布とする。

検定は、その分布に対して、検定統計量の位置から「そのような事態」が生じる「確率」としてp値を計算するのである。

MCA2010=2021では、1-2）の方法による正規分布近似（平面の場合はχ^2分布近似）を行い、それに対応した検定統計量の変換からp値を求めている。本稿では、1-1）の方法でリサンプリングを実行することによって、この組み合わせ論枠組みの考え方を確認すると同時に、その中での1-1）と1-2）のアプローチによる検定値（検定統計量とp値）の比較を行なってみた。

1-1）の方法は、非復元抽出によるリサンプリングを実行することになるが、実装上問題になるのは、その繰り返し数Rである。全体の標本数がNで部分集合の度数がnとすると、考えられる組み合わせ数は、${}_{N}C_{n}$となる。18-24の年齢群の度数は、n=93であり、総数N=1215であるので、

$$_{1215}C_{93}=\frac{1215!}{93!(1215-93)!}=1.709412\times10^{141}$$

という巨大な値となる。この数の抽出を行う場合を「厳密」版と呼ぶが（方法1-0）、これはコンピュータの処理時間を考えても現実的ではない。こうした場合、Rを1,000から10,000程度で実行する「厳密」版からのランダムサンプリング版が考えられる（モンテカルロ法）。このときの繰り返し数Rは、「99以上、999以下の無作為な並べ替えで十分であろう」といわれている（Rizzo 2008=2011:258）。

ここで、組み合わせ論的枠組みと呼んでいるのは、計算方法だけでみれば、計算機集約方法のリサンプリング技法が用いられるということであるが、検定の考え方についてネイマン＝ピアソン型のオーソドックスな検定論との違いが横たわっている。

- この検定では、分析対象の確率分布を想定する必要がない並べ替え検定というリサンプリング技法が用いられる。
- また、その手法で計算される*p*値が、確率分布モデルによるものではなく、数え上げと割合によるものが強調される[4)]。
- また、ネイマン＝ピアソンの*p*-値利用の伝統とは異なり、それを、帰無仮説を棄却するか否かの判定閾値に用いるではなく、典型性、同質性のレベルを示すものとして用いるという使い方である。その意味では、このアプローチはネイマン＝ピアソンの判定論ではなく、Fisherの*p*値の考え方（「その差異」は更に検討をすすめるべき差異として評価してよいか、を判断し研究をすすめる）に属している［柳川 2018:42］。

表5　伝統的検定手法と組み合わせ論手法の対応

問題状況	伝統的手法		組合せ論的枠組み	
	前提	手法	前提	手法
1群の平均値の差	正規性	t検定	なし	典型性検定
2群の平均値の差	正規性、等分散	t検定	なし	同質性検定

4. 実験

組み合わせ論による検定計算を、MCAのリザルトによって取得された座標値と追加変数のカテゴリとを接続したデータをもとに検討していく。

ここでは、組み合わせ論による3種の計算（実際は2つ）だけでなく、t-検定、Welchの順位和検定での検定統計量、*p*-値の比較も行う。3種というのは以下の三種である。

厳密検定（方法1-0）：全体の度数をMとして、注目している部分集合の度数をnとすると、M個の全体からn個の集合を抜き出す（サンプリング）する数は、

$$ {}_{M}C_{n} = \frac{M!}{N!\,(M-N)} $$

近似検定1（モンテカルロ近似）方法1-1：これは以下で例としてあつかう、18-24歳年齢群の93個を、全体の1215からサンプリングする組み合わせ数として計算すると $_{1215}C_{93} = 1.709412\ x10^{141}$ となり、非常に時間がかかる処理となる。そのために、組み合わせ数のモンテカルロ近似として、10,000という繰り返し数を用いた。

近似検定2（統計分布近似）方法1-2：そのため、実用となるのは、このモンテカルロ近似と、さらに、最初から正規分布なりχ^2分布を近似的にあてはめることで、統計量を評価するアプローチである。MCA2010=2021：116-121の式による「近似」はこうした確率モデル分布による近似である。

その数値の「違い」をどう生かすのか、ということになると、当然ながら、*p*値以外の要素を、検討しなくてはならない。そこで、どのような検討素材の利用が可能か、ということも問題になる。

分析実験

こうした違いを踏まえて、以下の実験をおこない、計算結果を比較した。比較する以下の通りである。なお、追加変数としたデータは「年齢：Age」、比較した統計量は以下の通り。検討した次元は、1軸を中心にしている。

なお、諸手法を評価するために必要となる各方法による計算結果は表6の通りである。

表６　伝統的検定手法と組み合わせ論手法の比較 / 検定統計量と p 値

name	cnt	p.cnt/R	SD	dim.1	dim.1/SD	wt	p.val.Z	p.val.t	dim.stat	dim.p
18-24	0	0.0000	0.0631	0.5893	9.3437	93	0.0000	0.0000	9.3429	0.0000
25-34	0	0.0000	0.0358	0.2722	7.6051	248	0.0000	0.0000	7.5911	0.0000
35-44	510	0.0051	0.0350	0.0890	2.5442	258	0.0055	0.0058	2.5452	0.0109
45-54	10130	0.1013	0.0422	-0.0538	-1.2740	191	0.1013	0.1021	-1.2791	0.2009
55-64	0	0.0000	0.0430	-0.3668	-8.5212	183	0.0000	0.0000	-7.7039	0.0000
65+	0	0.0000	0.0363	-0.2805	-7.7263	242	0.0000	0.0000	-8.5052	0.0000

cnt　並べ替え分布のうち、検定統計量より多くに位置する度数を数えたもの。

p.cnt/R　cnt を並べ替え分布の総数（R）で除して、割合として計算した p 値。

SD　並べ替え分布から計算した標準偏差 SD（内容的には、標準誤差 SE）

dim.1　検定対象の座標値（第 1 軸）。

dim.1/SD　標準正規分布を仮定したときの検定統計量。

wt　各年齢部分集合の度数。

p.val.Z　帰無分布を正規分布として評価した時の、検定統計量（dim.1/SD）の p 値。

p.val.t　同じく、t- 分布として評価した時の、検定統計量（dim.1/SD）の p 値。

dim.stat,dim.p MCA2010=2021 の第 5 章に記述されており、GDAtols:: dimtypicality function に実装されている帰無分布自体に正規分布近似を用いた統計検定量と p 値である。モンテカルロ近似で生成した値との差は、この近似の方法の違いによる。

なお、dim.p の値が p.cnt/R、p.val.Z、p.val.t の 2 倍になっているのは、典型性検定は全体の平均である原点とのずれを両側検定であることによる。

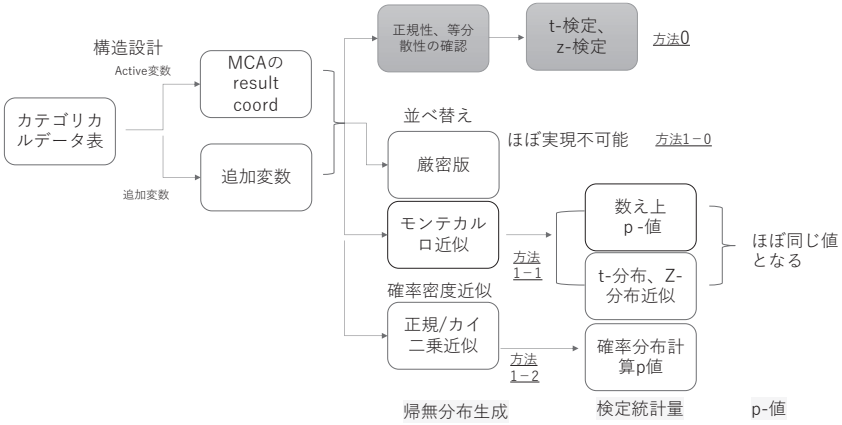

図 16　検定手法の関係

5. まとめ

IDA における典型性検定、同質性検定の計算アルゴリズムは、並べ替えである。この手法を用いることによって、伝統的統計学では必須となり、また、それゆえに、検定式の根拠としては曖昧さを残したままの分析となっていた正規性や等分散性の仮定を回避することが可能になっている。

もちろん、ここで典型性検定にしろ同質性検定にしろ、入手している標本データの内部での「偶然性」を検討するものでしかなく、母集団についての言明が可能になっているわけではない。母集団との関係については、標本の抽出において無作為抽出によって代表性が維持されているかどうか、また、十分な回答率が確保されることでその前提が維持されているかどうかを抜きに論じることはできない。

そうした意味で IDA で実施される検定は、検定内容を明確にした検定である。

こうした視点は、伝統的な検定論においても同様に求められるものであるにもかかわらず、記述統計と推測統計を、確率論で結びつけることによって、検定しているものがなんであるのかが曖昧になっている。

GDA の 1 ステップである IDA は、並べ替え検定というリサンプリング手法もさることながら、その計算過程を通じて、検定においてなにが検定されているのか、を具体的に明らかにすることを可能にしている。

この帰納的データ解析の方法としての意義はそこにもあると言える。

Appendix

A 並べ替え検定でのリサンプリングの考え方と、ブートストラッピングによる検定の実行

並べ替え検定は、全体データからのリサンプリングを行う。そこでは注目している部分集合とそれ以外という区分が行われ、注目している部分集合と同じ要素数が非復元抽出で行われる。そうやって抽出されたリサンプリング・サンプルの平均値の分布が**帰無分布**となる。

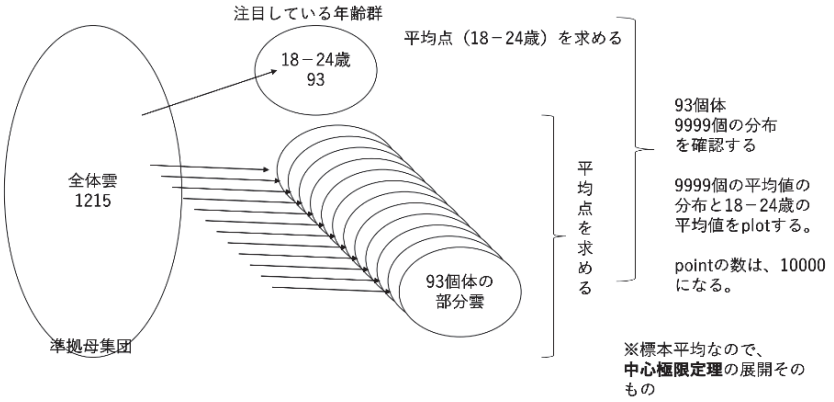

図 A1　並べ替えによる帰無分布の生成と中心極限定理

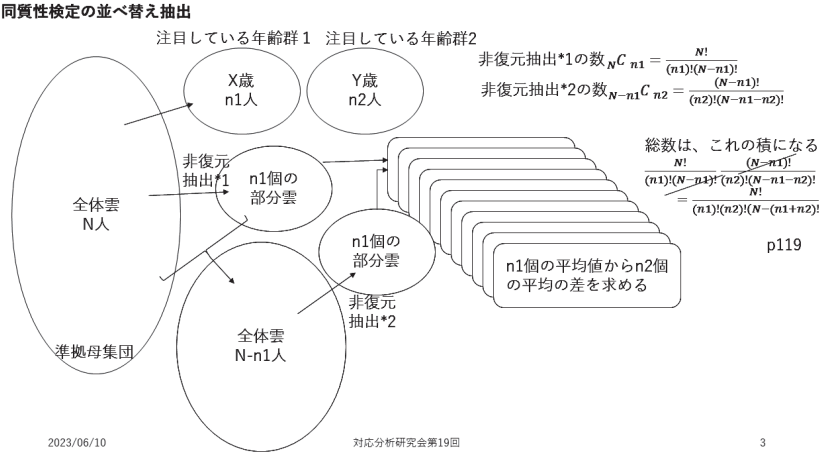

図 A2　同質性検定のための並べ替え抽出の手順

参考：ブートストラッピングによる検定

他方、ブートストラッピングは、全体データではなく、注目している部分集合がリサンプリングの対象となる。そこからのリサンプリングは並べ替えとはことなり復元抽出である。考え方としてはこうである。その標本が、全体から無作為抽出されたものであるとらえると、それをｎ倍すれば、母集団（に近いもの）が想定可能である。そこからの抽出であるので、復元抽出でよいということである。（サンプルリングされた数ｎ個すべてが、1番目のデータであることもありうる。並べ替えではそれはないので、非復元での抽出となる）

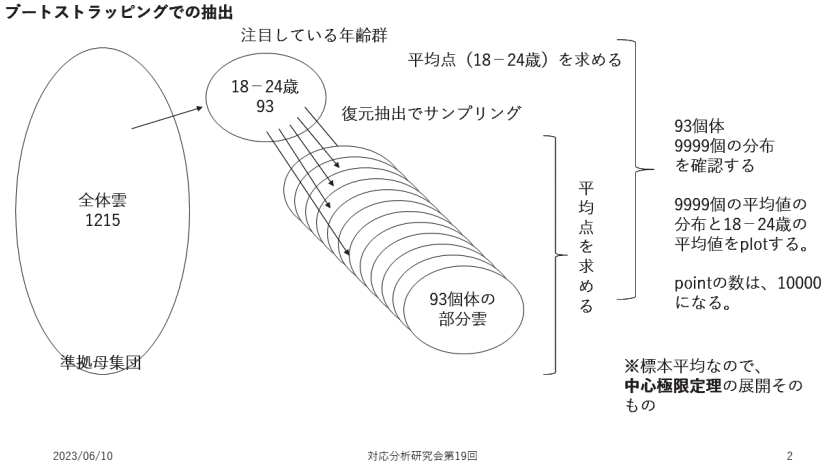

図 A3　参考：ブートストラッピングによる検定標本の抽出

18-24 歳の 93 個標本に対して、ブートストラッピングを行。
［Crawley 2015=2016 :91］

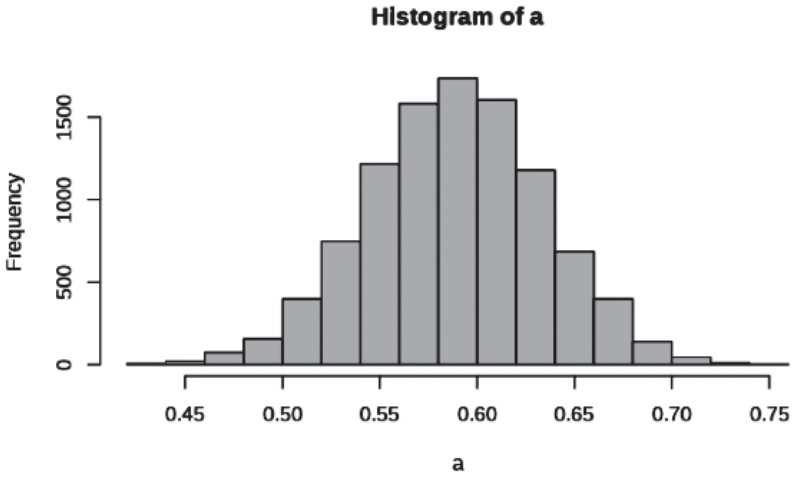

図 A4　参考：ブートストラッピングによる検定

謝辞

本稿の内容は、「対応分析研究会」（東京芸術大学、磯直樹先生主宰）での私の報告（とくに第19回、20回、21回）と、そこでの質疑応答に負っています。また、同内容を踏まえて、実際のデータ（「現在日本の文化と不平等に関する研究」）をもとにした、日本社会学会96回大会での報告と質疑応答に負っています。ともに、ご質問いただいた皆様に感謝いたします。もちろん、内容的な誤り、限界は藤本の責任であることはいうまでもありません。

なお、本研究は、科研費基盤研究（C）「データの幾何学的配置に着目したカテゴリカルデータ分析手法の研究」（20K02162）および、基盤研究（B）「現代日本の文化と不平等に関する社会学的研究：社会調査を通じた理論構築」（22H00913）の助成を受けています。記して感謝いたします。

注

1）このデータはLeRoux&Rouanet2010=2021で使われているデータで、LeRouxのWebサイト（https://helios2.mi.parisdescartes.fr/~lerb/Logiciels/software.html）から英語版をダウンロードすることができる。ここでは変数カテゴリを日本語化して用いている。

2）集中楕円（concentration ellipses）点の分布を楕円で近似表示する。点の分布が正規分布で均一であると仮定し描画される。この楕円に囲われた範囲には、全体の86.47％が含まれていることになる（Le Roux&Rouanet2010=2021:97、174）。

3）対応分析（CA）/多重対応分析（MCA）の「やっかい」なところは、一見わかりやすいグラフが提示されたとしても、その解釈に注意が必要なことである。平面に表示されたポイント間の関係を評価するにしても、生成された座標軸とそれへの寄与率を考慮する必要がある。1-2軸で近くに位置しているように表示されているポイントが、実は、3軸上では大きく離れているということもある。

そもそも、その近接性、離隔性自体が、比較困難な場合がある（CAの対称マップにおける行変数ポイントと列変数ポイントの距離をめぐる問題がある（藤本2017、Greenacre2017=2020:303を参照）。

多重対応分析の場合は、個体空間（CAの行変数空間）と変数空間（CAの列空間）を同時に表示することはしないので、この距離問題はあまり意識されることはない。

ただ、構造化データ解析によって、個体空間を、追加変数によって部分空間に分割して、その平均点を解釈する際には、平均点の座標関係だけではなく、その分散（の関係）に注目する必要がある。これは、初等統計学で、二つの群の比較を平均値だけで行うのは、不十分で分散も考慮する必要がある、とすることと同じである。

そこで、相関比η^2が分析において重要な意味をもつことになる。

SDA における個体分析では、1）で述べた集中楕円とこの相関比が重要な分析要素である。

4）一般的な統計学の教科書には、記述統計と推測統計を確率論がむすぶ、と書かれている。ここでいう確率論とは、データの生成モデルを確率モデルとして想定するということであり、それゆえに、推定論、検定論の際に、前提とする確率分布に適合していることの確認が必要になる。組み合わせ論的枠組みではそうした前提は必要としない、という意味で、「統計的推測を現在よりもより自由に用いることができるし、また、もちいるべきである」と書かれている。Le Roux & Rouanet 2010=2021:113

参考文献

Crawley, Michael J. , (野間口 , 謙太郎・菊池泰樹) 2016.『統計学 :R を用いた入門書 . 改訂第 2 版』. 共立出版 .

Freedman, David, Robert Pisani, Roger Purves. 2009. *Statistics*. Fourth edition, First Indian edition. Viva-Norton Student Edition. New Delhi Mumbai Chennai Kolkata Bengaluru Hyderabad Kochi Guwahati: Viva Books.

Hjellbrekke, Johs. 2019, *Mutliple Correspondence Analysis for the social Science*, Routledge.

Le Roux, Brigitte, Henry Rouanet. 2004. *Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis*. Dordrecht: Kluwer Academic Publishers.

Le Roux Brigitte; Henry Rouanet（大隅昇 ; 小野裕亮 ; 鳰真紀子）2010（2021）.『多重対応分析』. オーム社 .

Le Roux, Brigitte, Solène Bienaise, Jean-Luc Durand. 2019. *Combinatorial inference in geometric data analysis*. Chapman & Hall/CRC

Rizzo, Maria L, (石井一夫・村田真樹). 2011.『R による計算機統計学』. オーム社 .

Robette N.（2023）, *GDAtools : Geometric Data Analysis in R*, version 2.1, https://nicolas-robette.github.io/GDAtools/

東京大学教養学部統計学教室 . 1991.『統計学入門』. Tōkyō: 東京大学出版会 .

柳川堯 . 2018.『P 値 : その正しい理解と適用 .』統計スポットライトシリーズ 3 , 近代科学社 .

R Core Team（2023）. _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

藤本一男 . 2020.「対応分析は〈関係〉をどのように表現するのか」『津田塾大学紀要』51 号, pp156-167.

——— 2022.「日本における「対応分析」の需要を踏まえて、EDA（探索的データ解析）の中に対応分析を位置付け、新たなデータ解析のアプローチを実現する」『津田塾大学紀要』54 号, pp172-193.

——— 2023.「幾何学的データ解析（GDA）では分散はどのように分解されるのか ―GDA で ANOVA の手法を用いるために抑えるべきことがある―」『津田塾大学紀要』55 号, pp119-139.