様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成２５年　６月１８日現在

機関番号：８２６１６
研究種目：挑戦的萌芽研究
研究期間：2011～2012
課題番号：２３６５３２３３

研究課題名（和文）分布により与えられる多変量データの分析法に関する研究

研究課題名（英文）Multivariate analysis methods for frequency-valued data

研究代表者
宮埜　壽夫（HISAO MIYANO）
独立行政法人　大学入試センター　研究開発部・教授
研究者番号：９０２００１９６

研究成果の概要（和文）：心理学の研究においては、反応の変動性にもかかわらず、その平均のみを分析することが多い。本研究では、平均データではなく、反応の分布データを直接に取り扱う多変量解析法について検討した。とくに、区間値データに対する主成分分析、多次元尺度法における距離の計算、分割表における列変数の区間値による尺度化および列変数が複数である場合の尺度化について、混合分布モデルの考え方を利用する新たな方法を与えた。

研究成果の概要（英文）：In psychological studies, the mean values of responses are often treated as a main concern of statistical data analysis, neglecting the variations of responses. In this study, we addressed the issue of multivariate analysis methods for analyzing the frequency distribution data of responses, and, utilizing the idea of mixture distribution, developed a principal component analysis method for interval-valued data, and a scaling method of successive categories for contingency table. Also the latter method was extended to contingency table with more than two variables.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	1,100,000	330,000	1,430,000

研究分野：社会科学
科研費の分科・細目：心理学・実験心理学
キーワード：多変量解析、尺度化、対応分析、多重対応分析、区間値データ

１．研究開始当初の背景

心理学においては、個人の反応に変動があるにもかかわらずしばしば反応の平均を対象にした分析が行われる。分析結果の信頼性を評価するさいに、このことは大きな問題となる。反応の変動を直接に扱う統計的な分析法は、シンボリック・データ分析と呼ばれる一連の分析法を除けばほとんど知られておらず、心理学的なデータの分析には不十分な状態にある。

データ変動を考慮可能な新たな統計的分析の考え方および分析法の整備は、心理データの分析において最も重要な課題のひとつとなっている。

２．研究の目的

本研究は、区間値あるいはより一般的に分布として与えられるデータに対する統計的多変量解析について計量心理学の観点から考察することを目的とする。個人の反応データを扱う心理学においては、反応データの変動性にもかかわらず、反応の平均を求め、その結果を解析することが多く行われている。本研究は、平均データではなく分布で与えられるデータに対する分析法を構築し、心理学におけるデータ解析に新たな方法を提供しようとするものであり、具体的には、以下の２点を目的とする。

(1)心理学において多く利用される対応分

析法、主成分分析法、多次元尺度構成法などを取り上げ、多変量分布データに対応した方法を与えること。

(2)多変量分布データに対する分析法の一般的なアプローチの枠組みを与えること。

３．研究の方法

シンボリック・データ分析を参考に、まず、ヒストグラムにより与えられるデータの基本的統計量−平均、分散、共分散など−について検討するとともに、系列カテゴリー値をとる多変量のヒストグラムデータを分析する方法として、各カテゴリーを区間(1 次元解のとき)あるいは領域(多次元解のとき)として表現する新たな対応分析および多重対応分析を開発する。次に、類似性判断データにおいて対象間の類似度あるいは非類似度がヒストグラムにより与えられる場合について、多次元尺度構成法を開発する。

以上の研究結果に基づき、分布データに対する分析および分析法を開発する場合の統一的なアプローチの枠組みについて考察する。具体的には、以下について検討する。

(1) ヒストグラムにより与えられるデータの基本統計量について検討する。シンボリック・データ分析において仮定されているように、ヒストグラムの各区間においてデータは一様分布にしたがうことを仮定して平均、分散、共分散を定義する。

(2) 分布データの分析法を考えるために、系列カテゴリー値をとる変数に関する分割表および多重分割表の分析を取り上げ、カテゴリーが連続する区間（1 次元解）あるいは領域（多次元解）により表されるとして系列カテゴリーの尺度化および各行（例えば上の例では教科）の数量化を行う方法を開発する。さらに、開発された多重対応分析法について、主成分分析との関係を検討する。

(3) 分布データの分析法として、多次元尺度構成法について検討する。そして、これまでに検討した方法例から、分布データを分析するあるいは分析法を開発するさいの一般的な枠組みについて検討する。

４．研究成果

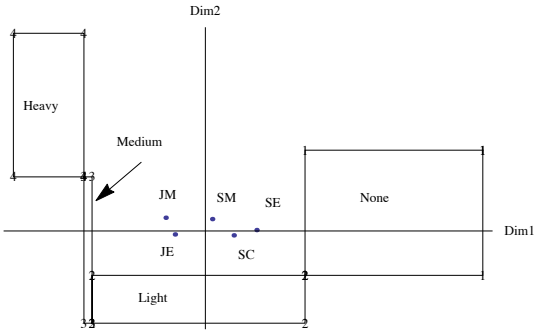
前節に述べた方法にしたがって研究を進めた結果、以下のような成果が得られた。

(1) 反応の頻度分布に対する分析法である対応分析およびその一般化である多重対応分析に対応する新たな方法が得られた。方法は、分割表あるいは複数の変数からなる分割表において、変数が系列カテゴリーである場合に有効な方法である。従来の分析法は、系列カテゴリーであることを考慮しない方法であり、そのために尺度値の信頼性が評価されなかったが、新たな方法はこの信頼性が評価可能な方法である。すなわち、系列カテゴリーを従来のようにひとつの点として尺度化するのではなく、連続する区間値（1 次元の場合）あるいは領域（多次元の場合）として尺度化する方法を新たに開発した。

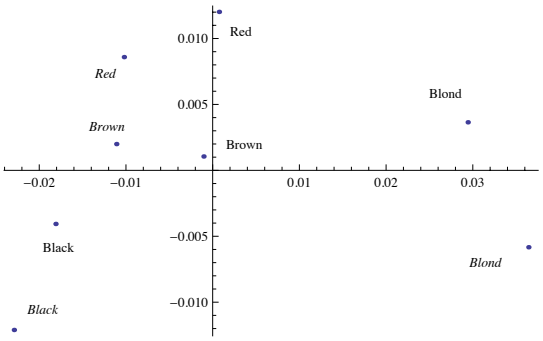
方法は、既存の対応分析と形式上類似しているが、イナーシャなどの対応分析における基本的な概念との関係などは明らかではない。また、複数の系列カテゴリー変数を扱う多重対応分析についても同様に形式上は類似しているが、関係はいくつかの点で明らかではない。既存の方法との関係を明らかにすることが残された課題である。

分割表の分析例として、喫煙習慣データおよび目の色と髪の色との関係データの分析結果を示す。喫煙習慣データは、喫煙の程度を 4 段階に分けて、職種ごとに喫煙者数を表したものである。下図(a)より明らかなように、中程度の喫煙カテゴリーの尺度値は他のカテゴリーの尺度値に比べ、変動の少ないことが分かる。

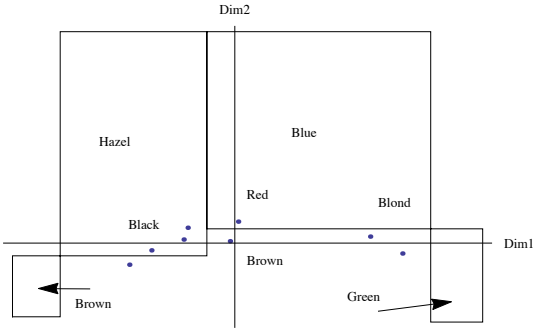
また、目の色と髪の色との関係データは、性別ごとにまとめられた目と髪の色の分割表データであり、目の色は Brown, Hazel, Blue, Green の 4 色、髪の色は Black, Brown, Red, Blond の 4 色である。下図(b-1)および(b-2)は分析結果であり、(b-1)は髪の色の尺度値であり、イタリックは女性を表している。また、(b-2)は Green の目と Blond の髪との関係などがすぐに分かる。



(a)喫煙習慣データの分析結果



(b-1)髪の色の尺度値
(イタリックは女性の場合を表す)



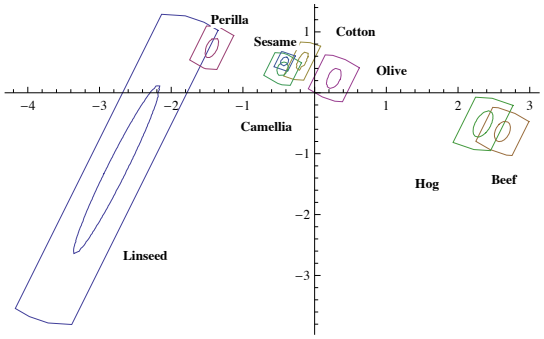
(b-2) 目の色と髪の色との関係
（髪の色は男性の位置のみ示してある）

(2) シンボリック・データ分析の代表的な多変量解析法の例として、主成分分析法が挙げられる。しかし、変数の増加にともなって処理対象となる行列の行数が飛躍的に増加すること、主成分得点の範囲が大きめに推定されることなどの問題点のあることが知られている。前節の(1)に述べたように、ヒストグラムの各区間においてデータは一様分布にしたがうことを仮定し、区間値データを特別な場合として含む一般的なヒストグラムデータに対する主成分分析法を開発した。すなわち、新たに開発された分析法は、ヒストグラムを生成するモデルとして、各区間の相対頻度を混合率とする一様分布の混合分布モデルを仮定した方法である。

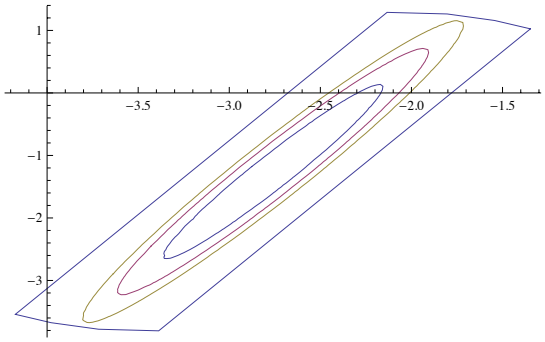
開発された方法は、従来の区間値データに対する分析法の拡張であると同時に、データ行列の行数の飛躍的増大、得点の変動範囲の過大推定の問題を解消した方法である。また、得点の変動範囲は、基本的に多次元空間における凸領域の主成分空間への射影であること、凸領域の射影は凸であること、および相対的内点は射影によって相対的内点であることに着目し、２次元の主成分空間における得点の厳密な変動範囲の推定方法を与えた。また、いくつかの区間値データを解析することにより、混合分布モデルによる方法は、この厳密な推定方法による解の良い近似であることを示した。

右図(a)は、よく知られた油脂成分データの分析結果である。楕円形で表された領域は、混合分布モデルによる主成分得点の推定された領域(確率約５０％の等確率楕円)、それを含む凸領域は厳密な主成分得点の領域を表している。また、右図(b)は、Linseed について確率を約５０％、８０％、９５％と変化させたときの等確率楕円と厳密な解領域との関係を表している。混合分布モデルに基づく解領域が厳密な解領域の良い近似であることが分かる。

なお、従来の方法による領域は厳密な凸領域を含む最小の矩形領域である。油脂成分データにおいては、Linseed の場合に最も過大に領域が推定されている。



(a) 油脂成分データの分析結果



(b) Linseed の等確率楕円と厳密解

(3) 分布データに基づく多次元尺度構成法については、対象間の非類似度を主成分分析の場合と同様なデータから定義する方法を主に検討した。その結果、同一対象内での変数間の相関を無相関と仮定するとき、自己非類似度がゼロとは異なる対称な非類似度行列が得られることを明らかにした。

自己非類似度がゼロでない、すなわち対角成分がゼロでない非類似度行列は、対象が何らかのカテゴリーである場合に自然に現れる行列であり、Tversky らによって多くの研究がなされている。このことから、非対称の非類似度モデル（距離−密度モデル）の特別な場合である「対象間の非類似度は、対象間距離と対象の自己密度との和と単調関係にある」ことを仮定したモデルによる分析法が妥当であると思われる。距離−密度モデルの推定方法はすでに知られているので、実際にデータを分析し、モデルの有効性を検討することが今後の課題である。

(4) 基本統計量に関する考察および主成分分析法、複数の系列カテゴリー変数を含む分割表の分析法などの開発の結果、分布により与えられる多変量データを分析するための一般的なアプローチとして、混合分布モデルを基本とする方法が有効であるとの結論が得られた。混合分布モデルは、数学的な扱いが容易であり、モデルの解釈も単純であるという利点を有している。

さらに多くの分析法に対して混合分布モ

デルによるアプローチを適用し、アプローチの妥当性を検討する必要がある。

なお、本研究を進めるにあたり、分割表の分析についてはニューカッスル大学（オーストラリア）の Eric Beh 准教授、主成分分析についてはライデン大学（オランダ）の Pieter Kroonenberg 教授の協力を得た。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計２件）
①宮埜　寿夫、荒井　清佳、シンボリック・データの主成分分析について、第 76 回日本心理学会大会、2012 年 9 月 11 日、専修大学
②宮埜　寿夫、統一試験の諸問題について、応用統計学会シンポジウム、2011 年 10 月 15 日、成蹊大学

６．研究組織
(1)研究代表者
宮埜　寿夫（HISAO MIYANO）
大学入試センター・研究開発部・教授
研究者番号：90200196

(2)研究分担者
荒井　清佳（SAYAKA ARAI）
大学入試センター・研究開発部・助教
研究者番号：00561036