様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３１　年　　４　月　２１　日現在

機関番号：　１７１０２
研究種目：　挑戦的萌芽研究
研究期間：　2016～2018
課題番号：　１６Ｋ１３４１７

研究課題名（和文）多変数順位相関係数の提唱と応用

研究課題名（英文）Formulation and Application of Multivariate Rank Correlation Coefficient

研究代表者
鈴木　譲（Suzuki, Yuzuru）

九州大学・人間環境学研究院・教授

研究者番号：４０２８１７５３

交付決定額（研究期間全体）：（直接経費）　　1,800,000円

研究成果の概要（和文）：　本研究の目的は、社会調査の質と精度の改善のために新たな計量分析手法を提案することである。社会調査における収集データは厳密には連続変量ではなく、離散変量である。離散変量を扱うクロス表での分析手法には順位相関係数があり、具体的な指標にはKendallの順位相関係数や、Goodman-Kruskalの順位相関係数などがある。しかしながら、これらはすべて扱える変数が２変数までという制約がある。
　本研究ではこの制約を取り除き、扱える変数を３変数以上の多変数に拡張し、多変数順位相関係数を構成した。そして実際の調査データを用いて分析を行えるように、マイクロソフト・エクセルの下で稼働するアドインを開発した。

研究成果の学術的意義や社会的意義
　本研究の学術的、社会的意義は大きく２点にまとめられる。第一に、社会調査のデータを連続変数ではなく離散変数として扱う場合に、扱える変数の個数の制限を取り除いたことである。これまでは、離散変数としての分析を行うには、順位相関係数を用いるのが通例であったが、この手法は２変数に限定されている。今回開発した多変数順位相関係数は、離散変数として３変数以上を扱うことができ、これまでのような制約はない。第二点目として、今回の分析手法においては、多重回帰分析のように変数間に従属変数、独立変数という非対称性を設定する必要がない。従って、因果関係が明確でないような一連の変数に対しての分析においては特に有効である。

研究成果の概要（英文）：　The goal of this research project is the formulation of a new analytical method for better quatlity and for higher precision of quantitative research.　In a strict sense, the data obtained in the social survey are discrete variables rather than continuous ones.　In the cross-tabulation method, technics for handling discrete variables do exist such as Kendall's rank correlation coefficient and Goodman-Kruskal's rank correlation coefficient.　These methods can handle, however, only two variables.
　In this project, we eliminated this barrier and developed what we call "Multivariate Rank Correlation Coefficient" or MRCC which can handle two or more variables.　For practical purposes, we also developed a software package which runs under Microsoft Excel.　The algorithm that we developed in this project has no limitations in terms of the number of variables.　From practical perspective, however, there are some constraints.　The software that we developed can handle up to 50 variables.

研究分野：　社会学

キーワード：　計量分析　離散変数　順位相関　多変量解析　社会調査　クロス表

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

社会調査における計量データの分析手法は、大きく次の２つに分類される。

・連続属性を持つ変数としての分析手法：多重回帰分析、因子分析など
・順序属性を持つ変数としての分析手法：クロス表分析、順位相関係数など

しかしながら、これまでの分析手法は、連続変数としての分析に極端に偏っていた。順序属性を持つ変数の分析手法は、1948 年に M.G. Kendall が順位相関係数を提唱し、その後、1954 年に L.A. Goodman & W.H. Kruskal がその修正を提案して以来、ほとんど進展を見ていない。現に近年出版されている社会調査のテキストにおいても、これらの手法を超える内容は何ら紹介されておらず、しかも、その適用は２変数のクロス表分析に限定されていた。

そこでまず、これまでの連続変数に関する既存の計量分析の手法を吟味し、たとえばロジスティック回帰分析を用いる際に確率概念がいかに曖昧なままに使用されているかを指摘した。しかしながら、より根本的な問題は、意識調査の回答など本来は非連続の順序属性を持つ変数を、無条件に連続変数として扱っていることであり、順序属性を前提する分析手法に立ち返る必要があるとの結論に至った。つまり、連続変数を前提とする数学的モデルから離れる必要があるということである。

２．研究の目的

本研究の目的は、社会調査の質と精度の改善のために新たな計量分析手法を提案することである。現状では、クロス表分析手法である順位相関係数には２変数の制約があるが、これを３変数以上の多変数に拡張し、多変数順位相関係数を構成することが目的である。

意識調査での回答者の評価値は、あくまで順序としての値であって、連続変数の取る値と見なすには明らかに無理がある。この点はこれまでもしばしば指摘されている。にもかかわらず、連続変数を用いた数学モデルが扱いやすいことと、クロス表の順位相関係数は２変数が前提のために３変数以上には直接適用できない、この２つの理由から、多変量解析は連続変数を前提に展開されてきた。しかしその結果として、数学的モデルを緻密化・複雑化しているほどには、意識調査のデータ解析の質と精度は向上していない。

また、連続変数を用いた多変量解析においても、多重回帰分析では従属変数と独立変数の区別が本質的であり、Pearson の相関係数のように各変数を同等に扱う相関関係の分析手法は、３変数以上には拡張されていない。本研究で提唱する多変数順位相関係数は、上記の２つの問題点を解決しようとするものである。

３．研究の方法

本研究では、順序属性を前提として、すべての変数を同等に扱い、多変数における順位相関を-1 から+1 の範囲の値で表すような指標を開発する。２変数の場合には、クロス表分析の手法としても用いることができるものを目指すが、本質的に、Kendall の τ、Goodman-Kruskal の γ のいずれとも異なる新しい指標を構成する。この指標の数学的構成方法の概要は、以下の通りである。

k を順序属性を持つ変数の個数、n をサンプルの大きさとする。変数を２変数から多変数に拡張する方法としては、直感的には２次のクロス表から k 次のクロス表への拡張が想定されるが、このやり方では適切な拡張を行うことができない。そのために本研究では n 行 k 列の行列を考える。各行がサンプルに対応し、k 個の変数の値が並んでいる。n 個のサンプルから、2 つのサンプルを取り出し、サンプル対を構成する。この作業をすべての組み合わせに関して行い、全部で nC2 だけの個数からなるサンプル対の集合を構成する。サンプル対の変数ごとに大小関係を調べ、この集合を以下の 4 つの部分集合に排他的に分類する。

P: 対の各変数の値がすべて異なり、大小関係がすべて揃っている
Q: 対の各変数の値がすべて異なり、大小関係が混在している
R: 対の変数の値に、同じ値と異なる値が混在している
Z: 対の各変数の値がすべて同じである

P, Q, R に含まれるサンプル対の個数を、それぞれ p, q, r とする。多変数相関係数 ω は、次の式によって定義される。　$\omega = (p - q)/(p + q + r)$

ここで、P が正相関、Q が負相関、R が無相関に対応し、Z は考察対象からは除外される。さらに、分析モデルに関するパラメータとしては、負相関に対する得点計算において各対に重みづけを与えるオプションや、同順位を持つ対を許容するオプションなどを考慮する。これらの扱いは、数学的には明確に定義できるがサンプル数が大きくなると手作業で計算するのは現実的ではないので、分析ソフトウエアを開発する。

４．研究成果

本研究で多変数順位相関係数を開発した際の根本的な問題は、多変数の場合に「順位相関」をどのように定義するかであった。すなわち、多変数の場合に相関の正、負、無相関をどのようにとらえるかである。本研究では、多変数における相関を「すべての変数が同時に規則的に変化すること」として定義した。

まず「規則的に変化する」ということの意味であるが、これは各変数の変化を個別に見た場合に、その変数の変化の方向が一貫していることと考えた。各変数の変化の方向の組み合わせは、方向をすべて反転した場合は重複しているからこれを除いて考えると、全体で 2^(k-1)だけある。従って規則的な変化といっても、2^(k-1)の規則性がある。言い換えれば、変化の方向の一貫性といっても 2^(k-1)だけの一貫性があることになる。つまり 2^(k-1)の異なった相関が共存しているわけである。

ここで、各変数を個別に見た場合に、その変数の変化の方向に一貫性があるかどうかは、1つのサンプル対ではたかだか 1 つの変化しか分からないから判断ができない。一貫性があるかどうかを見るためには、サンプル対が 2 つ以上必要であり、加えて、各サンプル対でその変数の値が変化している必要がある。当然、すべての変数について変化の一貫性を調べるためには、すべての変数において変数の値が変化している必要があるから、同順位を 1 つでも持つサンプル対は具体的な規則性を調べる対象にはならないことになる。

ただ、同順位を持つ対については 2 通りの場合を区別しなくてはならない。まず、同順位でない変数が 1 つ以上あるようなサンプル対である。この場合は、変化していない変数があるのに、変化している変数もある、という状況であるから、この対自体が、変数全体としての相関性はない、つまり無相関（相関 0）を示していると考えられる。次に、すべてが同順位の場合である。すべてが同順位の場合には、変数全体として相関があって値が無変化なのか、それとも、無相関であるがたまたま値が無変化なのかは区別がつかない。無相関であれば、いくつかの変数の値が無変化だとしても、残りの変数については値が無変化の場合もあれば変化する場合もあるからである。従ってこのような対は、変数全体としての相関、無相関に関して、何ら有益な情報を提供していないと考えられる。

以上の議論から、全体集合 T を以下の部分集合の直和として表現した。言い換えれば、以下の分類カテゴリーに排他的に分類した。ここで、Pi はそれぞれ異なる相関に対応している。

Pi　同順位が 1 つもない場合（1≦i≦2^(k-1)）

Ri　同順位が 1 つ以上あり、同順位でない変数も 1 つ以上ある場合
(1≦i≦((3^k)-(2^k)-1)/2)

R0 同順位でない変数が 1 つもない場合（すべて同順位の場合）

R0 は分類カテゴリーとしては単独であるが、多変数の相関、無相関に関して有益な情報を提供していないと考えられるから、考察から除外し、T から R0 の要素を除いた集合を S として、S を全体集合としてとらえることとした。あるいは、S における条件付き確率を考えると言っても良い。

なお、0≦i≦k-1 とする時、i 個の同順位の変数を持つような分類カテゴリーの個数は、次に示すように kCi・2^(k-i-1)である。まず k 個の変数の内、どの変数に i 個の同順位を持つ変数を割り当てるかは kCi だけの組み合わせがある。そしてそれぞれの場合に、残りの k-i 個の変数は同順位を持たないから、これら k-i 個の変数を 2 通りの大小関係により分類する方法は全部で 2^(k-i)だけある。ただ、大小関係を反転させた場合は同じ分類カテゴリーであるから、この重複を取り除くために 2^(k-i)を 2 で除した値、つまり 2^(k-i-1)を先に計算した kCi に乗じれば良い。これとは別に、i = k の場合、つまり、すべて同順位の場合は R0 に相当し、その分類カテゴリーの個数は 1 である。

同順位が 1 つもない場合は、i = 0 であるから、分類カテゴリーの個数は 2^(k-1)である。同順位が 1 つ以上あり、同順位でない変数も 1 つ以上ある場合は、i = 1 から k-1 までの和となるから、二項定理を用いて計算すると、((3^k)-(2^k)-1)/2 となる。すべての分類カテゴリーの総数は、i = 0 から k までの和で、やはり二項定理を用いて計算すると、((3^k)+1))/2 となる。

S が Ri と Pi の直和として表現されることは、無相関を示す対と、異なる相関を示す対の集合に排他的に分類されていることに相当する。各 Pi においては、変数の変化の規則性は一貫しており相関が認められるわけであるが、これはいわば局所的な規則性、相関であり、局所相関と呼ぶことができる。問題は S 全体として、いわば全域的、大域的にどのように相関を判断するかである。異なった局所相関が共存しているということは、いくつかの変数が同じように変化した場合、残りの変数の変化に関して、異なった規則性が混在していることを意味している。

無相関には、いくつかの変数が変化していないのに残りの変数が変化する場合と、いくつかの変数が同じように変化しているのに残りの変数の変化に規則性がない場合の 2 通りがある。前者は 1 つのサンプル対から判断できるが、後者は複数のサンプル対を用いて判断しなくてはならない。前者は Ri に対応し、後者が今回の判断に対応する。

局所的な規則性が異なった形で混在しているということは、全体として見れば規則性がない、つまり、無相関ということになる。従って大域的に相関をとらえるための考え方としては、異

なった局所相関の組み合わせを無相関として相殺し、結果としてどのような局所相関が残るのかを判断することになる。ここで問題は、相殺のためにどのような組み合わせを用いれば良いかを判断する基準がないことである。異なった局所相関は、どのような組み合わせを用いても、いくつかの変数の変化は同じで、残りの変数の変化の規則性が必ず異なっている。つまり、異なった局所相関は、どのような組み合わせを用いても何らかの形で相殺可能である。しかし、それではどのようにして異なった局所相関を相殺して、結果的にどの局所相関が残るのかを判断することができない。

そこで本研究では、どの局所相関に対して他の局所相関を相殺するかという基準は、分析者が指定する必要があると考えた。順序属性は値を反転させれば、変数の値の増減は入れ替わるから、順序属性の値を適切に設定して、すべての変数の変化が正となるような局所相関を基準として設定すれば良い。これは、座標軸を反転させると考えても同じである。このようにすれば、サンプル対の変数値の変化で正負の値が混在しているような局所相関は、すべて負相関として扱い、基準となる局所相関と相殺されると考えることができる。

以上の議論に基づき、基準となる局所相関を正相関とし、それ以外の局所相関を負相関として扱う。正相関のカテゴリーの対には 1 を付与し、負相関のカテゴリーの対には負の値を付与し、無相関のカテゴリーの対には 0 を付与して、重み付け平均、つまり、期待値を計算することとした。ここで、負相関のカテゴリーの対に一律に-1 を付与する方法も考えられるが、ここでは座標軸で基準相関のカテゴリー、つまり象限に近いものと遠いものとで付与する値を区別することにする。その理由は、基準となる座標軸は分析者が決めたもので固定されており、座標軸を自由に入れ替えられるわけではないから、これを基準にして象限の遠近を考えるのは意味があると判断したからである。

基準となる正相関のカテゴリーを（+, + . . . +）と表す。これは、(-, - . . . -)と同値であるから、正相関と負相関で符号が異なるのは、最大でも、k が偶数ならば k/2、k が奇数ならば(k-1)/2 である。符号の異なる座標が少ないほど、局所相関としてのカテゴリーの類似性が高く相殺効果は小さいと考え、逆に符号の異なる座標が多いほど局所相関としてのカテゴリーの類似性が薄く相殺効果は大きい、と判断する。具体的には、異なる符号の個数を u とする時、k が偶数であれば-2u/k を各対に付与し、k が奇数であれば-2u/(k-1)を各対に付与する。この値は、半分の符号が異なる場合に-1 となり、1 つしか符号が異ならない場合には-2/k または-2/(k-1)となる。

異なる符号を u 個持つ相関カテゴリーの個数は、kCu である。ただし、k が偶数で u = k/2 の場合には、相関カテゴリーの個数は kCu / 2 となる。これらのカテゴリーの要素数の合計を Qu とする。また、正相関のカテゴリーの要素数を P、無相関のカテゴリーRi の要素総数を R とする。k が偶数か奇数かに応じて、多変量順位相関係数ωをこれらの値を用いて二通りに定義した。具体的には、正相関のカテゴリー要素に 1、無相関のカテゴリー要素に 0、負相関のカテゴリー要素に-2u/k または-2u/(k-1)を対応させる確率変数の期待値として定義した。ωは-1 から 1 までの間の値を取る。

以上は同順位を持つ対を無相関として扱う方法であるが、これに加えて、条件を緩和して同順位を持つ対も相関カテゴリーに含めるアルゴリズムも開発した。具体的には、同順位でない変数について、すべて正の相関を示しているものは、正の相関カテゴリーに属する対として扱い、それ以外のものは負の相関カテゴリーに属している対として扱うこととした。ただ、同順位でない変数の個数を v とする時、正の相関としては v/k の値を付与し、負の相関としては k が偶数か奇数かに応じて-2u/k または 2u/(k-1)の値を付与することとした。

これらのアルゴリズムは、数学的には明確に定義されており問題はない。しかし、実際のデータ分析を手作業で計算するのは余りに煩雑であり非現実的である。そこで、マイクロソフト・エクセルの下で稼働するアドインを Visual Basic を用いて開発した。入力データとしては、連続した範囲にある n 行 k 列の順位データを扱うものとした。ここで、n はサンプルの大きさ、k は変数の個数である。各セルに入る値は順位を示す 1 から 100 までの自然数、サンプルの大きさは 3 以上 5000 以下、変数の個数は 2 以上 50 以下とした。

入力パラメータは、以下の通りである。

- データ範囲の指定
- データ範囲の先頭行がラベルか否かの指定（デフォルトはラベルなし）
- 重み付け：デフォルトか定数値かの指定
- 出力数値の精度：デフォルトは小数点以下 6 桁（指定値は 2 桁から 10 桁）

正常出力値は、以下の通りである。

- サンプル対の総数
- 分析から除外されたサンプル対の個数
- 分析対象となったサンプル対の個数
- Kendall と Goodman-Kruskal の順位相関係数（2 変数の場合のみ）
- 同順位を許容しない場合の多変数順位相関係数ω
- 同順位を許容しない場合の正相関、負相関、無相関の要素数
- 同順位を許容しない場合の正相関、負相関、無相関の得点数
- 同順位を許容する場合の多変数相関係数ω
- 同順位を許容する場合の正相関、負相関、無相関の要素数

・同順位を許容する場合の正相関、負相関、無相関の得点数

このソフトウエアを用いてデータ分析を行い、多重回帰分析などの従来の手法と分析結果を比較したが、事前に期待していたほどの顕著な対比は見られなかった。この意味で、今後とも引き続き、具体的なデータを用いて実証研究を続けて行く必要がある。今後の研究の展開のためには、従属変数を明確に定めることが難しく、かつ、連続変数として扱うのが難しいような離散変数のデータを厳選し分析を行うことが重要と考えられる。

また、今回の研究ではあくまで記述統計のレベルでの議論に焦点をあてていたが、今後の議論面での発展のために、推測統計への応用についても考察を行った。しかしながら、基本となる確率分布を適切に定めることができなかったので、この応用に関しては今回の研究期間内には確固たる結論を導くことはできなかった。

ただ、この推測統計の応用について検討している過程で、統計的仮説検定の手法の一つである比率の差の検定に関して、検定方法自体に問題があることが判明したので、この点については、論文において指摘し、学会報告も行った。また、データの平均性と中心性という二つの異なる性質が、多くの場合に混同されており、誤解を生じていることも判明したので、この点についても論文、学会報告において指摘を行った。

＜引用文献＞

① M.G. Kendall, 1948, *Rank Correlation Methods*, London: Griffin.
② L.A. Goodman and W.H. Kruskal, 1954, "Measures of Association for Cross Classifications," *Journal of American Statistical Association*, 49: 732-764.
③ 鈴木譲、ロジスティック回帰分析と確率、西日本社会学会年報、第 13 号、2015、99-113

５．主な発表論文等

〔雑誌論文〕（計３件）

① 鈴木譲、比率の差の検定における不合理、共生社会学、査読無、第 9 号、2019、1-8
② 鈴木譲、代表値の平均性と中心性、共生社会学、査読無、第８号、2018、1-11
③ 鈴木譲、多変数順位相関係数の構築、共生社会学、査読無、第 7 号、2016、1-10

〔学会発表〕（計６件）

①鈴木譲、比率の差の検定の有効性、第 91 回日本社会学会大会、2018 年
②鈴木譲、統計的仮説検定の不合理―比率の差の検定、第 76 回西日本社会学会大会、2018 年
③鈴木譲、データの平均性と中心性、第 90 回日本社会学会大会、2017 年
④鈴木譲、平均と分散―平均性と中心性、第 75 回西日本社会学会大会、2017 年
⑤鈴木譲、多変数順位相関係数の構築、第 89 回日本社会学会大会、2016 年
⑥鈴木譲、多変数順位相関係数の提唱、第 74 回西日本社会学会大会、2016 年

６．研究組織

(1)研究分担者　なし

(2)研究協力者　なし