分割表における多重比較の可視化ソフトウェアに関する研究

M2016SS012 安西祐輝

指導教員：松田眞一

1　はじめに

分割表のデータを分析する際には, 一般的に分割表全体に関して独立性に対する検定と対応分析が利用されており, 対応分析の結果は統計解析ソフト R でグラフ化することが可能である. 一方で, 松田 [3] は名義尺度の分割表に対する多重比較を提案しており, 分割表全体だけでなく分割表の内部まで分析することができる. 分割表について対応分析と松田 [3] が提案した多重比較を同時に利用する時, 対応分析のグラフに多重比較の結果を手作業でいれる方法であるが, これでは大きな手間がかかる. そこで, 本研究では, 分割表の対応分析と多重比較の出力結果を可視化するために, 統計解析ソフト R を用いて Excel 上で自動計算を行いグラフの出力までを行うソフトウェアを作成する.

2　本研究での分割表の定義

本研究では表 1 のような要因 X が a 個, 要因 Y が b 個ある $a \times b$ 分割表を利用する. ここで, 分割表のセル (i, j) において観測された度数を f_{ij} として, 第 i 行の度数の和を R_i, 第 j 列の度数の和を C_j, 全ての度数の和を N と定義する.

表 1　$a \times b$ 分割表

	Y_1	Y_2	$\cdots$	Y_j	$\cdots$	Y_b	計
X_1	f_{11}	f_{12}	$\cdots$	f_{1j}	$\cdots$	f_{1b}	R_1
X_2	f_{21}	f_{22}	$\cdots$	f_{2j}	$\cdots$	f_{2b}	R_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_i	f_{i1}	f_{i2}	$\cdots$	f_{ij}	$\cdots$	f_{ib}	R_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_a	f_{a1}	f_{a2}	$\cdots$	f_{aj}	$\cdots$	f_{ab}	R_a
計	C_1	C_2	$\cdots$	C_j	$\cdots$	C_b	N

また, 各セルの同時確率を p_{ij} とすると, 要因 X の確率分布は $p_{i\cdot} = \sum_{j=1}^{b} p_{ij}$, 要因 Y の確率分布は $p_{\cdot j} = \sum_{i=1}^{a} p_{ij}$ で表せる. ただし, $\sum_{i=1}^{a}\sum_{j=1}^{b} p_{ij} = 1$ である.

3　対応分析

対応分析は, はじめに行方向で基準化した分割表に対して, 固有値問題で数量化得点を算出した後, 行方向の数量化得点で双対性を使い列方向の数量化得点を求めるものである. (鄭・金 [7], 中村 [4] 参照)

本研究で作成するソフトウェアでの数量化得点 (図 1, 図 2 参照) は, R の出力グラフと同じにするため, 数量化得点に固有値の平方根で求められる正準相関係数を掛けた値が出力されるようになっている.

4　多重比較

$a \times b$ $(a \geq 3)$ 分割表に対して, 松田 [3] は閉検定手順を用いて以下の方法を提案しており, 本研究ではこの方法を用いて多重比較を行う. 以下にその要約を示す.

閉検定手順を利用するために, 最初に以下の帰無仮説の族 S を考える.

$$S = \{p_{ij}/p_{i\cdot} = p_{i'j}/p_{i'\cdot}\ (1 \leq j \leq b) \mid 1 \leq i \leq i' \leq a\}$$

すべての仮説が成り立つ場合は独立性の帰無仮説に一致する. さらに S を拡張して, いくつかの行の集まりでのみ構成される仮説の族を S' で表す.

$$S' = \{p_{i_1 j}/p_{i_1\cdot} = \cdots = p_{i_k j}/p_{i_k\cdot}\ (1 \leq j \leq b) \mid 1 \leq i_1 < i_2 < \cdots < i_k \leq a\}$$

ここでの k は仮説の大きさと呼ばれる. この S' に対して, 以下のように閉検定手順を適用する. 個々の仮説の検定に対しては Tukey-Welsch の方法と同様の有意水準をあてはめる.

手順 1　有意水準 α を定める.

手順 2　分割表全体に対して有意水準 α で独立性の検定を行う. 棄却されれば $k = a - 1$ とおいて次へ進み, 保留されれば S' のすべての仮説を保留して終了する.

手順 3　S' でまだ保留となっていない大きさ k のすべての仮説について, 対応する行のみを取り出した部分分割表に対する有意水準 α_k の独立性の検定を行う. ただし, α_k は式 (1), (2) で与えられる.

$$\alpha_{a-1} = \alpha \tag{1}$$

$$\alpha_k = 1 - (1 - k)^{k/a}\ (k = 2, \cdots, a - 2) \tag{2}$$

この独立性の検定で保留された大きさ k の仮説を含む (行の添え字の集合としては含まれる) 仮説は, すべて検定を行わずに保留する.

手順 4　大きさ k のすべての仮説が保留されている場合, または $k = 2$ の場合は終了する.

手順 5　S' 内の大きさ k の仮説のうち, 一つでも棄却されれば $k - 1$ を新しく k とおき, 手順 3 に戻ってこの手順を繰り返す.

手順 2, 手順 3 で用いられる独立性の検定は, Pearson の χ^2 検定と Fisher の正確確率検定のどちらでも良いが, 松田 [3] はどちらか一方に統一することを推奨している. 本研究では, 棚瀬 [6] が作成したプログラムを使用しており, Pearson の χ^2 検定を利用した行の要因に関する多重比較に限るという制限がある.

5　支援ソフトウェアについて

浅井 [1] が作成したパス解析を支援するソフトウェアと, 野口 [5] が作成したグラフィカルモデリングと SGS アルゴリズムのソフトウェアのプログラムをベースにして, ソフトウェアの作成を行う. また, 対応分析の計算には R の MASS パッケージにある corresp 関数を用いて, 多重比較の計算には棚瀬 [6] が作成した R 関数を用いる. 本研究で作成したソフトウェアの実行プロセスを以下に記述する.

1. ユーザが R と Excel で初期設定をする.
2. 分析をおこないたい分割表のデータをテキストファイルとして作成し所定の場所に保存したのち, Excel でデータのファイルのパスを指定する.
3. ユーザが対応分析の次元数と多重比較の有意水準の設定をする
4. R の実行命令文を作成した後, バッチコマンドで R を実行して, 対応分析と多重比較の計算をおこなう.
5. R の計算結果を R でテキストファイルに保存する.
6. 保存されたテキストファイルの有無を Excel で確認させる.
7. Excel 上に計算結果を出力させるために Sheet の初期化をおこなう.
8. Excel で対応分析と多重比較についての計算結果のテキストファイルを読み込み, 対応分析の結果は Sheet[対応分析] に, 多重比較の計算結果は Sheet[多重比較] に出力させる.
9. Excel の Sheet[plot 図] に対応分析と多重比較の計算結果に基づくグラフを作成する.

6　支援ソフトウェアの使用方法

本研究で作成したソフトウェアの利用手順を記述する.

6.1　初期設定

この支援ソフトウェアを利用するにあたり, R と Excel で以下の設定をする必要があるため, 分析前におこなう.

- R

  棚瀬 [6] が作成した R 関数を使用するために, gtools パッケージのダウンロード方法を以下に記述する.

  1. R を起動する
  2. メニューバーのパッケージにある「CRAN ミラーサイトの設定」の中の「Japan」を選択する.
  3. ワークスペースに install.packages(”gtools”) とコマンド入力をおこなう.
  4. ワークスペースに library(gtools) とコマンド入力をおこない, エラーがでなければ設定完了である.
  5. 設定完了後, R の「閉じる」ボタンを選択すると「作業スペースを保存しますか」の項目がでるが「はい」を選択して終了する.

- Excel

  支援ソフトウェアの分析フォームを立ち上げて, 以下の手順で R を Excel によって動かせるようにする.

  1. 分析フォームのコマンド [R 初期設定] を選択後, R-●.●.● フォルダを選択する. ただし, ● の所にはユーザが利用している R のバージョンが入る.
  2. bin フォルダを選択する.
  3. ユーザが使用している R が 32bit であれば i386 フォルダを選択して, 64bit であれば x64 フォルダを選択する.
  4. アプリケーション Rcmd(環境によっては Rcmd.exe) を選択して完了である.

6.2　分析設定

分析をおこなう前に対応分析の求める軸数の設定と多重比較の有意水準の設定をおこなう.

Sheet[Input] を開き, 対応分析の次元数と多重比較の有意水準の値を入れる. ただし, 本研究で作成したソフトウェアでは, 対応分析の次元数は 10 以下で, 多重比較は 0 以上 1 以下の値 (通常は 0.05) をいれる.

6.3　データの選択

分析をおこなう分割表データを支援ツールフォルダの● table 内にいれる. なお, データはテキスト形式の $a \times b (a \geq 3)$ 分割表で行, 列ともにラベルが必要である.

分析フォームの▼のボタンをクリックすると, 支援ツールフォルダの● table 内のテキストファイルが一覧できるようになっており, この中でユーザが分析したいデータを選択する. その後, コマンド [対応分析と多重比較] をクリックすると自動計算がおこなわれ, R の実行結果を支援ツールフォルダの● table 内の● R_Kekka に出力できる.

6.4　結果反映

分析フォームのコマンド [結果反映] を選択すると対応分析の結果は Sheet[対応分析] に, 多重比較の結果は Sheet[多重比較] に反映ができる.

対応分析は結果が分かるように, 行の要因の各水準が列の要因のどの水準の特徴として捉えられるかを色をつけて一目で見やすくした. 特徴の決定の仕方はユーザがはじめに設定する次元数で異なるため, 以下に記述する.

- 1 次元の場合
  1. 行の要因の各水準と列の要因の各水準の数量化得点を正と負で分ける
  2. 正同士, 負同士で同色をつける
- 2 次元以上の場合
  1. 列の要因の各水準に色をつける
  2. 行の要因の各水準ごとに, その数量化得点のベクトルと列の要因の各水準での数量化得点のベクトルの内積を利用して角度を求める
  3. 2. で求めた角度の中で, 一番角度が小さい列の要

因の水準をその行の水準の特徴と考え, 行の要因の水準の箇所に列の要因の水準と同じ色をつける

多重比較は, 表形式にして棄却されたものには 1, 棄却されなかったものには 0 で表示できるようにした (図 3 参照). また, 各セルにコメントを挿入しており, セルにポインタを合わせるとそのセルの水準の組み合わせが出力されるようになっている.

6.5 グラフの作成

分析フォームで, 行の要因を最前部にするならば行優先, 列の要因を最前部にするならば列優先にトグルボタンを押して設定する. 設定終了後, 分析フォームのコマンド [グラフ作成] を選択することでグラフを作成できる. グラフは, ＋の記号で原点を表し, ユーザが設定した軸について対応分析の結果で出力された数量化得点に基づいた水準の配置を行い, 多重比較で棄却された水準同士を線で結ぶ. 軸の設定方法は, Sheet[plot 図] のグラフ表示の箇所で, B23 セルに x 軸, B24 セルに y 軸の数値をいれる. グラフで表すときに, 行の要因と列の要因の区別がつくように列の要因には水準名の外枠を赤色で囲うようになっている (図 4, 図 5 参照).

対応分析の次元数を 3 次元以上に設定をしても, 行の要因の各水準の特徴は設定された次元数で色分けすることが可能ではあるが, グラフ作成において本ソフトウェアでは 3 次元以上のグラフ化をすることができない. そのため, グラフは, 対応分析の次元数を 3 次元以上に設定しても 2 次元でグラフを表す.

次に, グラフ表示に関しての工夫点を 3 点記述する.

- 多重比較の結果を見やすくするために, 行の要因の水準同士が重なったときには, 多重比較で棄却されたものを前部に配置できるようにしている.
- グラフが密着して, 要因の水準名が見づらいときに図の倍率を上げることで要因の水準名が見やすくなるようにした. 図の倍率の初期設定は 1 で, 図の倍率を上げるためには Sheet[plot 図] にある図の倍率の箇所 (A20 セル) に値をいれることで倍率を上げられる.
- 作図オプションを作り, ユーザが選択した要因のみの表示や, 選択した要因を最前部に表示できるようにしている.

7 データ解析例

2011 年 10 月に内閣府が行った国民生活に関する世論調査 [2] の中で, 老後は誰とどのように暮らすのがよいかという項目のデータで対応分析と多重比較をおこなう. 行は 8 つの水準で誰と暮らすかという要因であり, 列は 4 つの水準で都市規模の要因である. ただし, 紙面の都合上, 都市規模の分け方は割愛する.

対応分析の次元数は 3, 多重比較の有意水準 α は 0.05 で解析を行い, 行優先で作成したグラフを図 4 に示す. 対応分析の結果は図 1 と図 2, 多重比較の結果は図 3 である.

行	scores:		
8	1	2	3
息子(夫婦)と同居する	-0.24656	0.02622	-0.0119
息子(夫婦)の近くに住む	-0.11131	0.026199	0.037472
娘(夫婦)と同居する	-0.162	-0.05746	0.032233
娘(夫婦)の近くに住む	0.073917	0.022284	0.0137
どの子(夫婦)でもよいから同居する	-0.07915	-0.12413	-0.0441
どの子(夫婦)でもよいから近くに住む	0.047593	0.050687	-0.02073
子どもたちとは別に暮らす	0.09863	-0.01601	0.01411
わからない	0.124211	-0.00239	-0.03642

図 1 対応分析の行の結果

列	scores:		
4	1	2	3
大都市	0.150916	-0.04634	-0.01488
中都市	0.043435	0.043548	0.011228
小都市	-0.18463	0.001968	-0.028
町村	-0.14963	-0.06943	0.05222

図 2 対応分析の列の結果

図 3 多重比較の結果

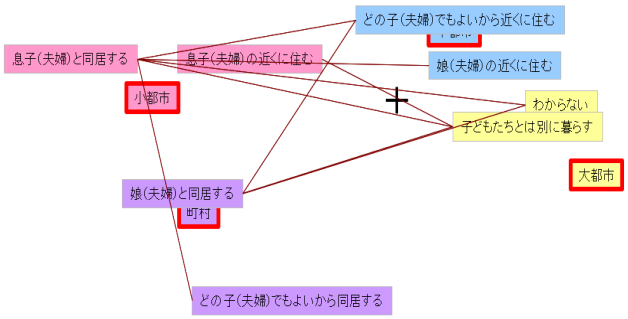

図 4 対応分析と多重比較の結果 (有意水準 0.05)

図 3 でそれを対応分析内でグラフ化した図 4 から老後に暮らしたい人と各都市間に以下の違いがあることが分かる.

- 小都市 – 息子 (夫婦) と同居する
- 大都市と中都市 – 子供たちとは別に暮らす, わからない, どの子 (夫婦) でもよいから近くに住む
- 小都市と町村 – 娘 (夫婦) と同居する
- 大都市と中都市と町村 – 娘（夫婦）の近くに住む
- 中都市と小都市と町村 – 息子（夫婦）の近くに住む

- 大都市と中都市と町村 – どの子 (夫婦) でもよいから同居する

行の水準である「息子 (夫婦) の近くに住む」は, 図 4 を見ると小都市には近いが, 共通の特徴である中都市や町村とは離れているように見える. しかし, 6.5 節より, グラフは 2 次元で表しているため, 中都市と町村から離れて見えると考えられる. また, その他の水準も同様に考えられる.

7.1 結果の考察

現在の結果だけでは, 共通の特徴が多く, 単独での特徴が少ないことから, 考察をおこなえるように明確な特徴が分かるようにする. はじめに, 1 軸, 2 軸, 3 軸ですべてのパターンの 2 次元のグラフを表したが, 明確な特徴が今回のデータからはわからなかった. 次に, 有意水準 0.1 にして解析を行うと特徴が見られたため, この方法で考察をおこなう（図 5 参照).

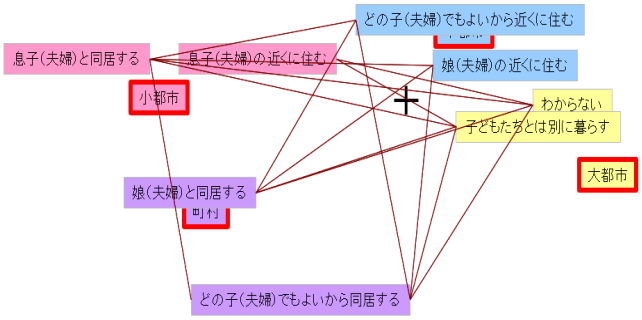

図 5 対応分析と多重比較の結果 (有意水準 0.1)

図 4 と図 5 を比べると, 有意水準 0.05 で棄却されず, 有意水準 0.1 で棄却された水準が見られた. 有意水準 0.05 で棄却されず, 有意水準 0.1 で棄却された理由は, この水準に回答する人が少なく, 有意水準 0.05 では証拠不十分であると計算されたからであると考える.

図 5 の結果を読み取ると, 大都市と中都市, 小都市と町村の 2 つの群に分けられることがわかる.

7.1.1 大都市と中都市

都会の人は子どもと別に暮らしたくて, 住んでも近くまでであることがわかる. 特に, 「別に暮らす」は小都市と町村の特徴のすべてで棄却されているため, 都会の人は別居をしたいと考えている. 別に暮らしたい理由として, 都会の人は, 人に関心をあまりもたないからであると考えられる. 都会の人は田舎に比べると地域密着も少なく, 人のつながりが弱いため, 自分の子供でも別々に暮らしたいという特徴がでたと考えられる. また, 都市は地価が高いため, マンションに住んでいる人も多く, 同居する部屋の確保が難しいことから別々に暮らしたい人が多いとも考えられる.

息子は小都市との共通の特徴にはなるが, 近くに住みたいと考える人が多いのは, 近くの考え方が都市と田舎で違うからである. 都市は, 交通機関が充実しており, 電車を使えば長距離も移動できるが, 田舎は交通機関が充実しておらず, 移動手段が限られることから距離を短く考えているかもしれない. そのため, 近くに住むは大都市や中都市の特徴としてでたと考えられる.

7.1.2 小都市と町村

田舎に住む人は, 同居したいと考え, 同居できなくても息子の近くには住みたいと考えている. このように考える理由として, 自営業の人であれば仕事を継いでもらいたいと思うことや, 都会に対して, 都会は人が多いため怖いと感じている人が多くいるからが挙げられる. 都会には自営業をおこなう人は多くないが, 田舎には農家などの自営業の人が都会に比べると多くいるため, 自分の代で仕事を終わらせないように子どもにも継いでほしいと考えるからである. また, 田舎の人は, 人のつながりを大切にするため, 自分の子供が都会に行き, 犯罪に巻き込まれないようにするために, 子どもと同居したいと考えていると思われる.

小都市の特徴である「息子と同居する」と町村の特徴である「どの子でも同居する」が棄却された理由は, 過疎化による影響だと考える. 町村は近年高齢化が進んでおり, 人口が減っている. そのため, 息子・娘のどちらでもよいから一緒に住みたいと思う人が多くいるからと考えられる.

8 おわりに

ユーザの作業を必要最低限にして, Excel 上のコマンドボタンを押すだけで, R の起動から計算をおこないグラフ化までを自動でおこなえるソフトウェアを作成したことから, R を使えない人でも利用できる. 本ソフトウェアでは, 行の要因に関する多重比較に限る制限をしたが, 棚瀬 [6] が作成したプログラムの改良をおこない, 列方向のみや行方向と列方向を同時に多重比較をおこなえるソフトウェアも作成することができると考える.

参考文献

[1] 浅井悟史: 従業員満足の因果分析に関する研究, 『南山大学紀要「アカデミア」数理情報編』, **12**, 45-55, 2012.

[2] 国民生活に関する世論調査: `https://survey.gov-online.go.jp/index.html`, 2017/12 閲覧

[3] 松田眞一: 名義尺度の分割表に対する多重比較法, 『南山大学紀要「アカデミア」数理情報編』, **4**, 29-37, 2004.

[4] 中村永友: 『R で学ぶデータサイエンス 2 多次元データ解析法』, 共立出版, 2009.

[5] 野口良輔: SGS アルゴリズムとグラフィカルモデリングに関する一考察, 『南山大学紀要「アカデミア」理工学編』, **15**, 13-21, 2015.

[6] 棚瀬貴紀: 『分割表における多重比較法とその評価』, 南山大学大学院数理情報研究科修士論文, 2006.

[7] 鄭躍軍, 金明哲: 『R で学ぶデータサイエンス 17 社会調査データ解析』, 共立出版, 2011.