論文・研究レポート

コレスポンデンス分析における布置の精度

齋藤　朗宏，豊田　秀樹

1. はじめに

コレスポンデンス分析は特に日本，フランス，米国等で頻繁に用いられている分析手法であり，マーケティング等の分野でも，ブランドイメージの分析といった場面でしばしば利用される手法である．この分析手法の主たる目的は，分割表に関して，行と列のそれぞれのカテゴリに数値を付与し，その同時布置により関係性をグラフィカルに表現することである．

一般的に，多くの統計解析手法では推定量の精度の評価について強い関心が持たれる．当然ながら，コレスポンデンス分析についても同様の関心を持たれてしかるべきであろう．文献[1]では，布置に表現される点の信頼領域に関しての考察がある．しかし，その結果は事象が厳密に多項分布に従うという厳格な仮定のもとでしか妥当といえない方法[2]であり，実際場面では精度の評価をされることはない．どの程度安定しているのかというのがわからない状態で，イメージの距離，関係などについて語るということには問題がある．

そこで，本研究では，POSの大規模データとリサンプリング手法を用いてコレスポンデンス分析の重みの信頼領域を考察し，コレスポンデンス分析の布置の精度と安定のための条件に関して考察する．

なお，コレスポンデンス分析は，理論上，多次元分割表で表現されたデータを多次元空間内に表現するものであるが，本研究では最も一般的に用いられており，しかも最も直感的に理解しやすい2次元分割表を2次元平面上に布置するものとする．

さいとう　あきひろ
早稲田大学 大学院文学研究科
〒162-8644 新宿区戸山1-24-1
とよだ　ひでき
早稲田大学 文学部
〒162-8644 新宿区戸山1-24-1
受付03.7.22　採択04.1.15

2. 方法

2.1 分析の方針

まず，本研究の理論的背景に関して説明する．一般に，推定量の精度に関して評価する方法は3通りが挙げられる．第1に，最尤推定のような分布に基づいた推定量を用いることで標準誤差を調べる方法がある．第2に，統計モデルの確率分布に基づいてシミュレーションデータを繰り返し発生させて調べる方法がある．第3に，リサンプリングを用いて推定量の分布を調べる方法がある．

コレスポンデンス分析は，分布に基づいた推定を行う手法ではないため，第1の方法を取ることはできない．また，シミュレーションデータを発生させる場合にはまず母集団全体を用いて算出した推定値，いわば真の構造を設定しなくてはならない．コレスポンデンス分析は，例えば分割表で対角要素にほとんどのサンプルが集中している場合のように，行と列の対応関係が綺麗に出ているときほど安定して同じような変数間の同時確率が見られる可能性が考えられ，真の構造そのものが精度に影響を与える可能性がある．その対策としていくつか真の構造を考えるにしても，変数間の関係には様々なバリエーションが考えられ，そのどれが実状とあっているのか，何が精度に影響を与えているのかを確認するのは困難である．

これらの理由から，精度に影響を与える要因が明確になっていない現時点では，できる限り実データで分析することが望ましく，第2の方法は適切とは言えない．そこで，本研究では第3の方法，つまりリサンプリングにより推定値を大量に発生させ，その分布を調べるという方法を採用する．

ブートストラップ法に代表される一般的なリサンプリング手法では，真の構造が不明である状態で分析を行わなくてはならない．しかし，もし極めて大規模なデータが利用可能であるならば，擬似的な母集団が存在しているとみなせる．それは真の構造があらかじめ

わかっている状態で分析できるということを意味している．そしてそこから非復元抽出を行うことには，母集団全体から繰り返しデータを取り，精度を評価する過程を正確に再現することができるというメリットがある．また，サンプリングする標本サイズを変えて比較する場合にも，同一の真の構造を持つ母集団から抽出したサンプル同士で比較することができるというメリットがある．

次に，コレスポンデンス分析において，精度に影響を与えると予想される要因を考える．まず，サンプリングする標本数を多くすることで推定精度が上がることが予想される．また，母集団においてある特定のカテゴリにサンプルが多く集中しているならば，そのカテゴリは安定して推定できるということも予想される．このうち，実験者が制御できるのは前者であるため，サンプリングの標本数をいくつか変えて分析を行う．また，前者は実質的にコレスポンデンス分析における大数の法則の確認とも言えるため，単一の母集団からの知見を一般化してもあまり問題がないが，後者に関しては，予想通りの結果となったとしても，他の分割表を分析した場合と比較しなければ一般化できない．そこで，比較用に分割表をもう1枚準備して分析を行う．

2.2 方法（分析1）

分析には，平成14年度データ解析コンペティションより提供された，2001年の某百貨店売上POSデータを用いた．ここから，分析1では化粧品関係のブランドのみの購入データを抽出し，その上で顧客の個人情報から年齢を抽出して購買データと結合し，年齢をカテゴリ化した上で分割表の形式にした．年齢によるカテゴリの分け方，分割表は表1に示されている．この分割表の特徴としては，①ファンケルハウスと資生堂が強く，あとの3ブランドはほぼ横並び，②25歳未満が圧倒的に多く，以下，年齢が上がるごとに人数は減少，③合計数が極端に違うので，分割表だけからでは年代と化粧品のブランドとの対応関係が見えない，の3点が挙げられる．③のように分割表だけからでは対応関係が見づらい場合に，コレスポンデンス分析は有効である．①，②の特徴に関しては後程考察する．

このデータに対して，まず全データを用いたコレスポンデンス分析で真の構造を得た．なお，本研究では，コレスポンデンス分析においてχ^2距離を主座標に基準化している．この結果，2次元での累積寄与率は99.19%となり，2次元で充分に説明されることが確認できた．次に，1000件，2000件，5000件，10000件のオブザベーションを非復元抽出し，その抽出されたデータセットに対してコレスポンデンス分析を実行するという手順をそれぞれ5000回繰り返した．その結果，それぞれ5000個の推定値が求められるので，それぞれに対して文献[3]に従って真の構造をターゲットとした直交プロクラステス回転を行った上で，その平均，標準偏差，縦軸と横軸の相関を算出した．最後に，算出されたデータを式(1)に代入し，95%確率楕円を描きそれを95%信頼領域の目安とする．

$$\frac{1}{2(1-\rho_{xy}^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}-\frac{2\rho_{xy}(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]=2.99573 \quad (1)$$

ここで，μ_xは全データを用いた場合の横軸の平均値，μ_yは同様に縦軸，σ_x, σ_yはそれぞれ横軸，縦軸の標準偏差，ρ_{xy}は縦軸と横軸の相関係数である．

2.3 方法（分析2）

分析2では，分析1と同じく2001年の某百貨店売上POSデータより菓子のブランドのみの購入データを抽出し，あとは分析1と同様の手順を踏んだ．その分割表は表2に示されている．

表1 分析1に用いた顧客人数

	～24歳	25歳～39歳	40歳～54歳	55歳～69歳	70歳～	合計
マックスファクター	1028	278	321	278	88	1993
ファンケルハウス	6244	1629	1167	678	185	9903
コーセー	1328	392	316	235	74	2345
鐘紡	1219	226	243	357	209	2254
資生堂	5171	763	785	1120	644	8483
合計	14990	3288	2832	2668	1200	24978

表2　分析2に用いた顧客人数

	～39歳	40歳～59歳	60歳～	合計
モロゾフ	3941	11320	11604	26865
ユーハイム	1348	6045	6868	14261
ヨックモック	3078	10395	9061	22534
合計	8367	27760	27533	63660

この分割表の特徴としては，④ユーハイムが一番少なく，モロゾフが多い[1]．⑤40歳未満がかなり少ない．⑥カテゴリごとの合計数がかなり違い，分割表からでは年代と菓子のブランドとの関係について対応関係を見ることができない，の3点が挙げられる．⑥は分析1の③と同様であり，④，⑤に関しては①，②同様後程考察する．なお，3×3の分割表のため，2次元までしか計算できず累積寄与率は100%となる．

分析2は，分析1との比較を目的とするため，10000件抽出，5000回繰り返しのみを行った．

3. 結果と考察

結果の布置に当たっては，分割表の行要素と列要素で別々に示す．以下にその理由を簡単に説明する．コレスポンデンス分析においては，文献[4]にも示されているように，行要素内，列要素内で布置に表現される点の間の距離はユークリッド距離としては評価できるため，その間隔を測ることに意味はある．一方で，行要素と列要素の間では，布置に表現される点の距離はユークリッド距離として評価できないため，間隔を測ることに意味はない．また，行要素と列要素の間でほぼ同位置に表現されるということは，カテゴリ同士に強い対応関係があるということを意味し，それ自体が重要な知見となり，信頼領域が重なったとしても問題ない．これらの理由から，精度の評価という観点からは，行要素と列要素との同時布置に意味がないため，本研究では同時布置を行わなかった．

3.1 結果と考察（分析1）

分割表における行要素，つまり化粧品のブランドに関する1000件，2000件，5000件，10000件抽出した場合の信頼領域はそれぞれ図1～4のようになった．

全体を通じて，ファンケルハウスと資生堂に関して精度が高いという特徴が見て取れる．この2ブランドは，方法において①として述べたようにカテゴリに反応したサンプル数が多い（つまり購入した顧客の数が多い）カテゴリである．その他のカテゴリに関しても分割表と比較すると，反応数と楕円の大きさとに対応関係があることが見て取れる．この点は，反応数が多いカテゴリ程精度が高いという仮説を支持していると

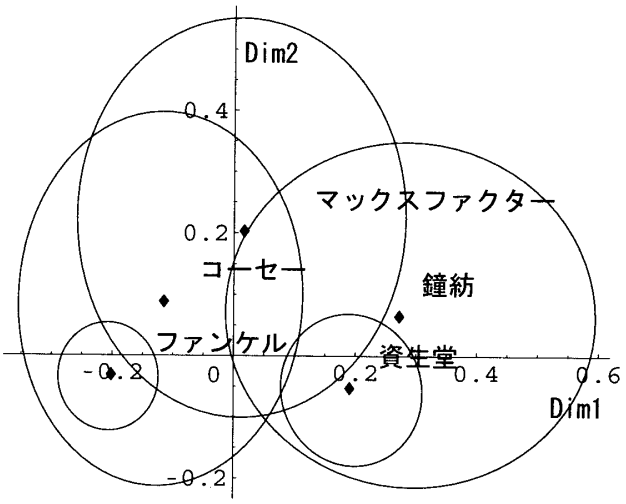

図1　1000件抽出した場合の行要素の信頼領域

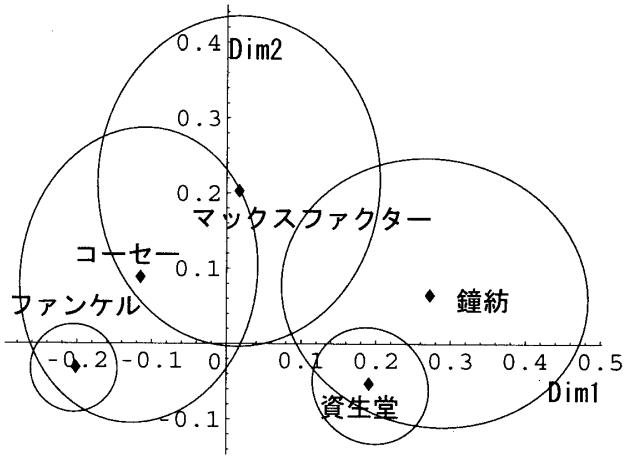

図2　2000件抽出した場合の行要素の信頼領域

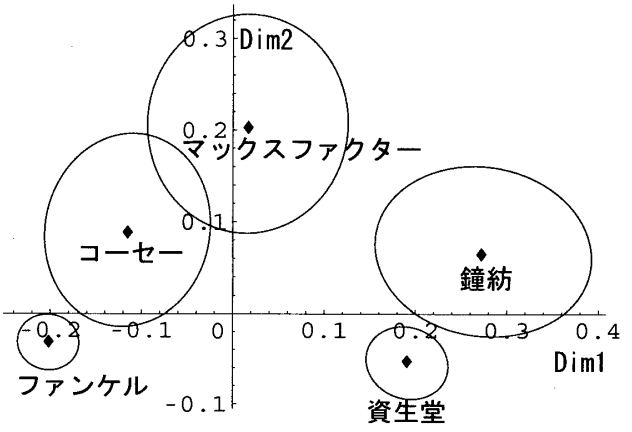

図3　5000件抽出した場合の行要素の信頼領域

[1] 件数にして，およそ倍の違いがあった．他2ブランドと比較するとユーハイムの少なさが目立つ．

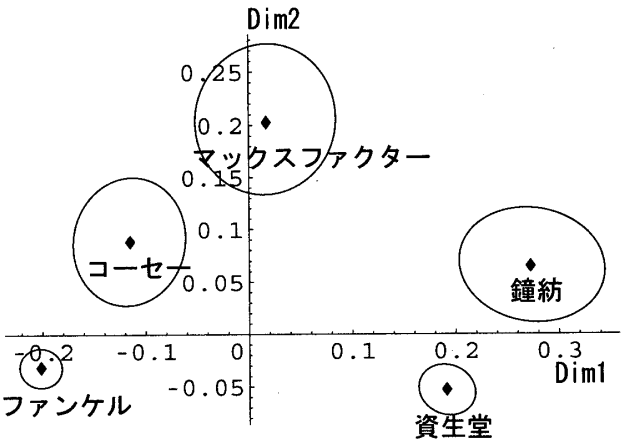

図4　10000件抽出した場合の行要素の信頼領域

考えられる．

また，サンプルサイズを増やすにつれ，楕円の形そのものは変化せず，他のカテゴリとの相対的な大小関係も変化しない状態で，信頼領域が徐々に小さくなっていることが確認できる．この点は，サンプルサイズを大きくすると精度が上がるという仮説を支持していると考えられる．

2変量正規分布の相関に関しては，紙幅の都合上図示はしないが，回転前には原点を中心に放射状になるような値を示していたが，回転後にはほぼ無相関となった．この結果は，以下すべての分析において同様である．

最後に，サンプルサイズと精度の関係が確認できたので，信頼領域の大きさから必要なサンプルサイズを考える．1000件や2000件では，特に反応数の少ないカテゴリに関して結果が全く信用できないことが確認できる．5000件～10000件は最低でも必要であると言えるだろう．

一方，分割表の列要素．つまり顧客の年代に関する1000件，2000件，5000件，10000件抽出した場合の信頼領域はそれぞれ図5～8のようになった．

すべてのサンプルサイズにおいて，25歳未満というカテゴリの精度が最も高くなっている．25歳未満は方法において②で述べたように，カテゴリに反応したサンプル数の多いカテゴリである．また，他のカテゴリに関しても，反応したサンプル数と楕円の大きさとに対応関係があることが確認できる．

また，やはり行要素と同様に，サンプルサイズを増やすにつれ，楕円の形はあまり変化せず，他のカテゴリとの相対的な大小関係も変化しない状態で，信頼領域が徐々に小さくなっていることが確認できる．

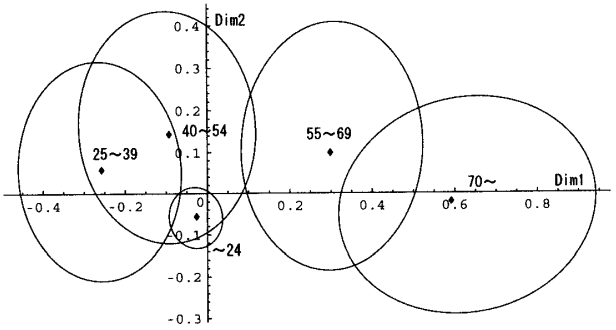

図5　1000件抽出した場合の列要素の信頼領域

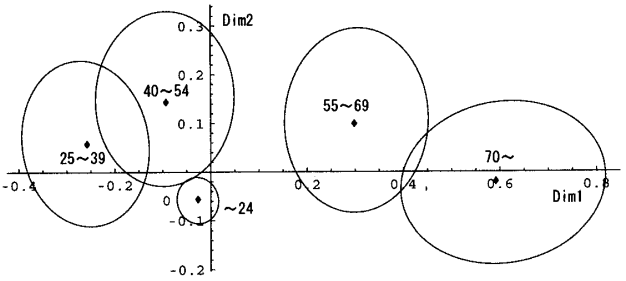

図6　2000件抽出した場合の列要素の信頼領域

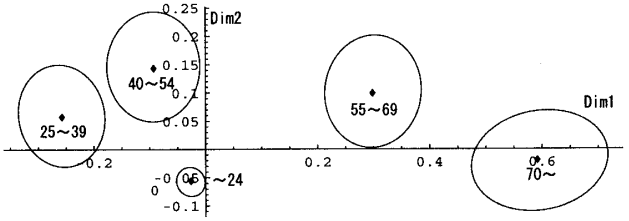

図7　5000件抽出した場合の列要素の信頼領域

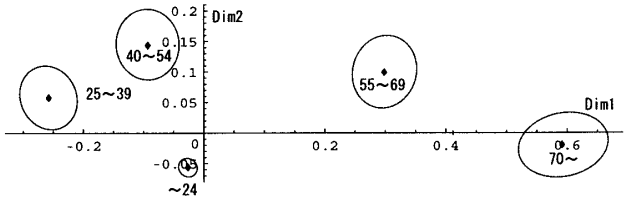

図8　10000件抽出した場合の列要素の信頼領域

最後に，サンプルサイズと信頼領域から，必要なサンプルサイズについて考察する．列要素の場合には，2000件から5000件の間で信頼領域が重なることがなくなり，分析が安定していることが見て取れる．行要素との兼ね合いで考えると，この変数群に関してコレスポンデンス分析を行う場合には，5000～10000件のサンプルサイズが必要であると考えられる．

3.2　結果と考察（分析2）

10000件を抽出してコレスポンデンス分析を行った場合の行要素，つまり菓子のブランドに関しての信頼

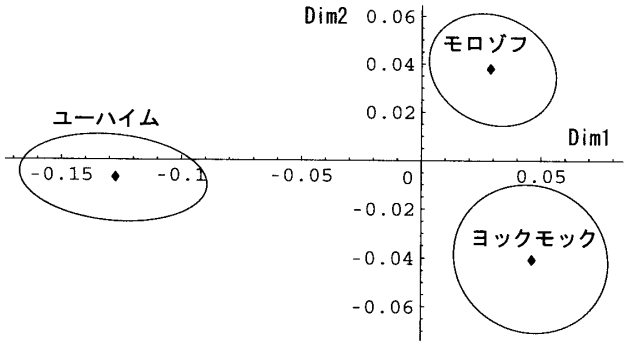

図9　10000件抽出した場合の行要素の信頼領域

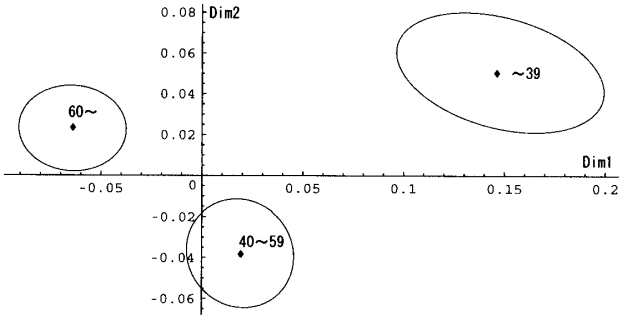

図10　10000件抽出した場合の列要素の信頼領域

領域は図9のようになり，列要素，つまり年代に関しての信頼領域は図10のようになった．

行要素，つまり菓子のブランドに関して信頼領域の大きさを比較すると，モロゾフが最も小さく，ユーハイムとヨックモックに関しては，形の違いもあって比較は難しいが，あまり違いがないと言える結果となった．反応数が多いカテゴリ程精度が高いという仮説に従うならば，モロゾフ，ヨックモック，ユーハイムの順で，ユーハイムは特に大きくなると考えられ，この仮説とは反する結果と言える．

一方，列要素，つまり年代に関して信頼領域の大きさを見ると，方法において⑤として述べた，反応するサンプル数の少ない40歳未満が最も大きくなり，反応数が多いカテゴリ程精度が高いという仮説を支持する結果と考えられる．

また，両図から，10000件のサンプルサイズがあれば，この変数群に関する分析は安定するということが確認された．

4．総合考察

本分析において確認された事項は，大きく分けて2点に分類できる．

第1点は，サンプルサイズは精度に直接的な影響を与えるということである．このことに関しては，分析1において，行要素，列要素の双方でサンプルサイズの増大とともに信頼領域が縮小した点からも明白であろう．

本研究における特筆すべき知見は，コレスポンデンス分析の精度という立場からは，5×5という非常に小さな分割表に関して分析を行う場合においてさえも，5000～10000という非常に大きなサンプルサイズが必要であるということが確認されたことである．実際には，26000件から10000件を非復元抽出という形を取っているため，精度に関して過剰によく評価されている可能性も否定できない．しかし，過剰に精度よく推定してもこの結果という点は，コレスポンデンス分析の不安定さに関して警鐘を鳴らすという意味で，本分析の意義を失わせるものではない．コレスポンデンス分析は，例えば今回のPOSデータを用いた分析のように，大サンプルが確保できる状況においてこそ安定した推定が可能であり，数百のサンプルしか準備できないような分野で行うには危険な分析法であることが示された．

第2点は，カテゴリに対する反応数は当該カテゴリの精度に強い影響を持つが，絶対的な要因ではないということである．このことは，分析1や分析2の列要素においてカテゴリに対する反応数と当該カテゴリの信頼領域の大きさとに対応関係が見られたのに対して，分析2の行要素に関して一部対応関係が見られなかった点から示された．本論文の結果からだけでは言い切れないが，行要素と列要素との対応関係が精度に関して無視できない要因として存在していることが示唆されたと言えるだろう．このような問題点を発見できたという意味で，実データからのリサンプリングを行ったことには意味があったと考えられる．

また，今後の課題として，上に挙げた行要素と列要素の対応関係の精度への影響や，ブートストラップ法のような復元抽出による精度検討の可能性についての研究，今回得られた知見のシミュレーションデータによる確認等が考えられる．

謝辞　本研究にあたりまして，慶應義塾大学の飯田孝久先生，レフェリーの諸先生には貴重なコメントを頂きました．この場をお借りして感謝申し上げます．

付録

```
/*-------------------------------------------
  リサンプリングを行い，コレスポンデンス
  分析の誤差を検討するSASマクロ
  著者　：斎藤朗宏（早稲田大学大学院）
  mailto: saito@suou.waseda.jp
  -------------------------------------------
  使用法:
  データセットは，2変数として作成する．例えば
  2 1
  3 2
  2 2
  （以下略）
  このとき，3 4は，変数群1に関して変数3に反応，
  変数群2に関して変数4に反応とみなされる．
  ===========================================
  %inc'このファイルの名前';
  %efa(
       fil = 'データファイル名',
       row=行データ（1番目の変数）の変数群名．,
       rows=行データのカテゴリ数 ,
       clm=列データ（2番目の変数）の変数群名．,
       clms=列データのカテゴリ数 ,
       size=1度のサンプリングで抽出する数,
       rep=サンプリングの繰り返し数,
       out=アウトプット
       );
  run;
  ===========================================
  たとえば，
  ===========================================
  %inc 'D:\sampmac.txt';
  %corresp(fil='D:\data.dat',
            row=age,rows=5,clm=kesho,clms=5,
            size=1000,rep=5000,out='D:\out.dat');
  ===========================================
  このプログラムは，1番目の変数群がageで5変数．
  2番目の変数群がkeshoで5変数．全データ中から，
  1000個のデータを5000回リサンプリングする．
-------------------------------------------*/
%macro corresp(fil=,row=,rows=,clm=,clms=,size=,
rep=,out=);%let _=_;
%let Dim1=Dim1;
%let Dim2=Dim2;
data alldata;infile &fil;
input &row &clm;run;
%resampling(size=&size,rep=&rep)

data count;%do reps=1 %to &rows+&clms+1;
counter=&reps;put counter 3.;output;
%end;run;
%do reps=1 %to &rep %by 1;
proc corresp data=samp&reps dim=2
outc=out&reps noprint;tables &row,&clm;run;

data dim1&reps;merge count out&reps;
if counter=1 then delete;
Dimall=Dim1;keep Dimall;run;

data dim2&reps;
merge count out&reps;
if counter=1 then delete;
Dimall=Dim2;keep Dimall;run;

data all&reps;set dim1&reps dim2&reps;run;
proc transpose data=all&reps out=trs&reps;run;
%if &reps=1 %then %do;

data allsamp;set trs&reps;run;%end;
%else %do;data allsamp;
set allsamp trs&reps;run;%end;%end;

data crspdat;set allsamp;file &out lrecl=100000;
put %do names=1 %to &rows+&clms+&rows+&clms;
col&names %end;;
%do names=1 %to &rows;
rename col&names=&row&_&names&_&Dim1;%end;
%do names=&rows +1 %to &rows+&clms;
%let namenum=%eval(&names-&rows);
rename col&names=&clm&_&namenum&_&Dim1;%end;
%do names=&rows+&clms +1 %to &rows+&clms+&rows;
%let namenum=%eval(&names-&rows-&clms);
rename col&names=&row&_&namenum&_&Dim2;%end;
%do names=&rows+&clms+&rows +1 %to
&rows+&clms+&rows+&clms;
%let namenum=%eval(&names-&rows-&clms-&rows);
rename col&names=&clm&_&namenum&_&Dim2;%end;run;
proc univariate data=allsamp;run;%mend corresp;

%macro resampling(size=,rep=);
%do reps=1 %to &rep;%let seed=%eval(152*&reps);
proc surveyselect data=alldata method=srs n=&size
out=samp&reps noprint seed=&seed;run;%end;
%mend resampling;
```

参考文献

[1] Benzecri, J. P.: "L'Analyse des Donnees (Tome 2), L'Analyse des Correspondances", Dunod, 1973.

[2] Hawkins, D. M.（医学統計研究会訳）：多変量解析の理論と実際, MPC, 1988.

[3] 芝祐順：因子分析法（第2版）, 東京大学出版会, 1979.

[4] *SAS/STAT User's Guide, Version 6 Fourth Edition*, Volume 1, SAS Institute Inc., 1990.