シミュレーションによる基データの特性を再現した簡便なデータ生成法

井上　正隆[1]　山田　　覚[2]

(2006年10月31日受付，2007年１月15日受理)

Simply and easy method of data generation which can reproduce data characteristic of basic ones by the simulation

Masataka Inoue　Satoru Yamada

(Received：October 31. 2006,　Accepted：January 15. 2007)

要　旨

人間にかかわるデータが正規分布に従うとことは，一般的に知られている。正規分布に従うデータの再現には，中心極限定理を利用できるが，質問紙調査のような複数の質問項目を含むデータを再現するためには，それら項目間の関係性も考慮する必要がある。本研究では，各質問項目の再現に中心極限定理を，項目間の関係性の再現には回帰分析を用いて，簡便なデータの生成法を開発した。

具体的なデータを用いて本簡便法を検証した結果，基データの各質問項目が持つ平均と標準偏差の特性が生成データに反映されていること，および各質問項目間の関係性が反映されていることが確認された。

本簡便法を用いることによって，倫理的な理由等により多数のデータが収集できない場合でも，大規模データを必要とする共分散構造分析などの分析が可能になると考えられる。

キーワード：シミュレーション・データ生成・共分散構造分析

Abstract

It is said that the data of the human being depend on normal distribution. While generation of data which depend on normal distribution can be reproduced using Central Limit Theorem, relationship among items of a questionnaire should be considered when data are reproduced with a questionnaire of some items. The simply and easy method of data generation was developed using Central Limit Theorem for reproducing each items, and using Regression Analysis for relationship among items.

As the result through real data, the method is able to reproduce the data reflecting characteristics of basic data and satisfy relationship among items. Using this method, it will be able to do Central Limit Theorem with large scale data and analyze small size data due to ethical consideration

Key word: simulation, data generation, covariance structure analysis

1 高知女子大学看護学部看護学科　助手　看護学修士　Department of Nursing, Faculty of Nursing, Kochi Women's University
2 高知女子大学看護学部看護学科　教授　工学博士　Department of Nursing, Faculty of Nursing, Kochi Women's University

1．はじめに

医療や看護サービスのoutcomeに関する研究を行う場合，その評価者はサービス消費者である患者やその家族が適当である。しかし，患者やその家族の調査協力は，医療従事者に比して得難く，分析に必要なデータ数を十分に確保できない状況にある。例えば，最近注目を浴びている，患者の受療行動の解析等に用いられている構造解析（共分散構造分析）をする場合では，各変数間の関係性をより複雑に分析するため，より多くのデータが必要となる（狩野ら2002，坂元1985，田部井2001，山田2002，山本ら1999，涌井ら2003)。

このような場合に，基データが持つ個々の変数の特性，およびそれら変数間の関係性を再現したデータ生成ができれば，採取したデータが少数でも，上記のような構造解析が可能となる。シミュレーションによりデータを生成し，分析に用いる方法は幾つか紹介されているが（有沢ら1997，薦田ら1995，中西1969，奥田1971)，その生成法は極めて複雑であり，日常的な看護や医療分野における研究に用いることは難しい。

本稿では，基データの持つ個々の変数の特性および変数間の関係性を，シミュレーションを用いて簡便に生成する方法を提案する。また，生成データが基データのデータ特性をどの程度再現できているか等を検証するとともに，本法の活用法についても検討する。

2．シミュレーションによるデータ生成の目的

ある概念を構成する要因と要因間の関係性を明らかにすることが研究の目的を果たすために不可欠な場合，これまで多く用いられてきた因子分析では，因子を抽出することはできるが，因子間の関係性や抽出された因子からさらに抽象度の高い潜在因子を抽出することはできなかった。一方，共分散構造分析ではこれらが可能であり，共分散構造分析を用いる方が，実際の現象に近い複雑な分析結果を得ることができ，より深く考察を加えることが可能な場合もある（狩野ら2002，坂元1985，田部井2001，山田2002，山本ら1999，涌井ら2003)。

しかし，共分散構造分析は，坂本（坂本1985）が「推定するパラメータの数が$2\sqrt{n}$を超えると，推定した値は，不安定な結果となり，望ましい結果は，得られにくい。」と指摘するように，用いるサンプル数に分析結果の安定性が強く依存する。このため，分析に用いるデータ数が十分に確保できない場合は，共分散構造分析を用いることをあきらめざるを得ない。

このような場合に，シミュレーションにより基データの特性を再現させたデータを生成することができれば，多数のパラメーターを用いた広域な範囲の分析を行えるようになる。

3．シミュレーションを用いた生成データに必要な条件

本稿で数値例として用いた生成の対象となるデータは，受療者行動等の人間にかかわるデータであることから，数値データは正規分布に従うと考えられる。再現すべきデータが正規分布に従うと仮定すると，平均と分散を生成することによりデータが再現できる。シミュレーションによる平均と分散の生成は、中心極限定理により行なえることが，一般的に知られている。

次に，シミュレーションを用いて生成したデータは，基データのデータセット（1枚の質問紙における全ての質問項目のデータで，個々の研究協力者や施設，ケースなどから得られたデータベースの最小単位をいう）内で変数間が持つ関係性を保持したものでなければ，正確な分析が行なえない。

また，生成されるデータは，分析に用いる分析手法に耐えうるだけのデータ数が確保されなければならないと考えられる。

以上のことから，生成されるデータは，以下に示す条件を満たさなければならない。

1）基データが，正規分布に従うこと。

2）基データの各質問項目が持つ平均と標準偏差の特性が，生成されたデータに反映されてい

ること。

3）基データの各質問項目間の持つ関係性が，生成されたデータに反映していること。

4）多数の件数を持つ広範囲なモデル検証が行える十分なデータ数を生成できること。

4．本簡便法のデータ生成方法

本法でシミュレーションを用いてデータを生成するために，シミュレーション分野の中で一般的に用いられている中心極限定理を用いて，基データよりデータを生成した。

中心極限定理とは，ある変数が0から1の区間において一様に分布する場合に，期待値すなわち平均が1/2に，分散が1/12になるというものである。よって，データ数(n)を12としてこの定理を用いると，以下のように基データの特性を持たせたデータを生成することができる。

手順1　正規乱数の発生

ある質問項目に対応した，すなわちその項目の基データの特性を反映した正規乱数x_iを発生させるためには，数式①にその質問項目の基データより算出された平均m_xと標準偏差s_xを代入することで算出できる（有沢ら1997，薦田ら1995，中西1969，奥田1971）。

ある質問項目の正規乱数　x_i

$=m_x+s_x\times\{(乱数_1+乱数_2+\cdots+乱数_{12})-6\}$

……数式①

m＝基データの平均

s＝基データの標準偏差

この作業をそれぞれの質問項目で行うと，質問紙内のすべての質問項目の基データより，それぞれの質問項目に対応した正規乱数を発生させることができ，1つのデータセットが生成できる（有沢ら1997，薦田ら1995，中西1969，奥田1971）。

しかし，質問項目間の関係性を考慮せず，各質問項目において正規乱数を発生させた場合，生成されたデータは質問項目間の関係性を持たない。そこで，データセット内で，質問項目間の関係性を反映した正規乱数を発生させる必要がある。

手順2　回帰式によるデータセット内での質問項目間の関係性の確保

質問項目間のデータ分布に関係性があるとすると，その関係性は回帰式で説明することができる。ある基準となる質問項目γを独立変数とし，別のある質問項目x_iを従属変数としての単回帰式を算出する作業を繰り返し，それぞれの変数をγとの単回帰式で説明できるようにする（一般的な記号の使い方が異なるが，ここでは求める変数をx_iで統一している）。基準となる質問項目γは，質問項目の内容が質的に他の質問項目の基準となるものを選出した。

ある質問項目x_i

＝基準となる質問項目　$\gamma_j\times a+b$

……数式②

a＝回帰係数

b＝定数項

また，基準となる質問項目γ_jと別のある質問項目x_kの単回帰式が成立しなかった場合は，質問項目γ_l（基準となる質問項目γ_jと単回帰式が成立した質問項目）と質問項目x_kで単回帰式を作成し，基準となる質問項目γ_jと質問項目x_kのデータ分布の関係性を質問項目γ_lを介して間接的に説明するようにする。

これによって，基準となる質問項目γ_jの基データの平均から，他のそれぞれの変数の平均を直接的か間接的に算出（補整）することができる。

手順3　生成するシミュレーションデータのデータセット内での関係性の確保

既述のように，データセット内でのデータ分布の関係性を保持して正規乱数を発生させるため，ある質問項目xに対応した正規乱数x_iを発生さ

せる際に用いる質問項目xの基データの平均m_xと，基準となる項目の正規乱数γ_{ji}（基準となる質問項目γ_jより生成した正規乱数）との関係性を保証する必要がある。基準となる項目の正規乱数γ_{ji}を基に，先に求めた回帰式（数式②）で質問項目xの平均m_xを補整し，質問項目xに対応した正規乱数x_iを生成する。（数式③）

質問項目xに対応した正規乱数x_i
$=(\gamma_{ji}\times a+b)+s_x\times(\Sigma$乱数$_1$＋乱数$_2$＋…
＋乱数$_{12}-6$）……数式③

a＝質問項目xを従属変数とし質問項目γ_jを独立変数とした回帰式の回帰係数
b＝質問項目xを従属変数とし質問項目γ_jを独立変数とした回帰式の定数項
s_x＝基データの標準偏差

手順4　正規乱数の基データに対応させた数値化

既述の手順に従って生成した正規乱数は，小数であり，例えば5段階のリッカート尺度を用いた調査の場合は，この得点に対応させるため，生成した正規乱数を整数化する必要がある。

基本的には，生成された正規乱数を四捨五入して整数化する。生成されたデータには，基データの5段階のリッカート尺度には存在しない－2未満と＋2より大きい数値が含まれる場合がある。よって，基データで最高点を＋2点，最低点を－2点としている場合には，－2未満は－2点とし，＋2点より大きい場合は＋2点として整数化する。

5．生成データが，基データのデータ特性をどの程度再現しているかの検証

総数23項目から成る著者らが行った研究で使用した基データ（n＝88）と本簡便法で生成したデータ（n＝25,000）の平均と標準偏差を図1と2にそれぞれ示す。どちらの図も基データを示す点線と生成データを示す破線がほぼ重なり，生成したデータと基データの平均と標準偏差は，視覚的に比較するとよく似ている。さらに，数値的に再現されているか調べる為，生成データと基データの間で，平均と標準偏差の適合度の検定を行なった。この結果，平均と標準偏差ともに生成されたデータは，基データに有意水準5％で適合しており，本簡便法は，基データの特徴（平均と標準偏差）を再現していることがわかった。

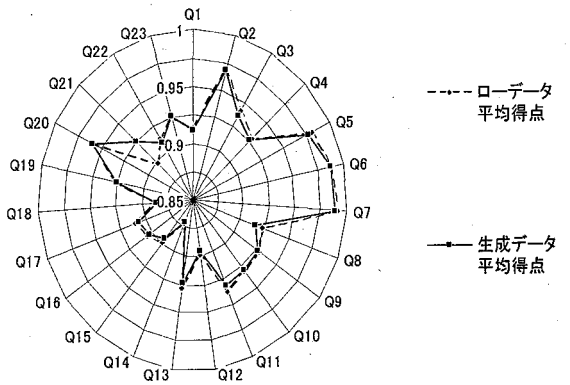

図1．生成したデータとローデータの適合度の検定（平均値）
χ^2値＝1.089×10^{-3}
χ^2値(棄却限界値)＝32.7(5%水準)

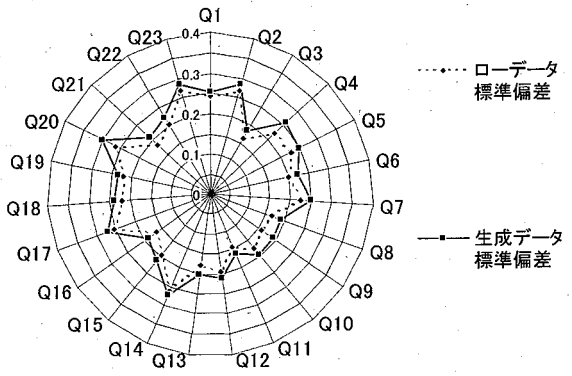

図2．生成したデータとローデータの適合度の検定（標準偏差）
χ^2値＝5.06×10^{-3}
χ^2値(棄却限界値)＝32.7(5%水準)

次に，本簡便法により生成したデータが，基データの変数間の関係性をどの程度再現しているかを評価した。筆者らが行った既述の研究で用いた任意の質問項目を用いてパス図を作成し，生成データと基データの両者を用いて共分散構造分析で他母集団の同時検定を行い，質問項目間の関係性を示す両データのパラメータ間の差を検証した。（図3，表1）

生成したデータと基データにおいて，それぞれの変数間の関係性が異なるのであれば，生成データと基データのパラメータ間に統計上有意な差が認められるはずであるが，他母集団の同時検定の

結果，表１に示すように生成データと基データのパラメータ間に統計上有意な差は認められず，生成したデータが，基データの変数間の関係性を再現していると言える。

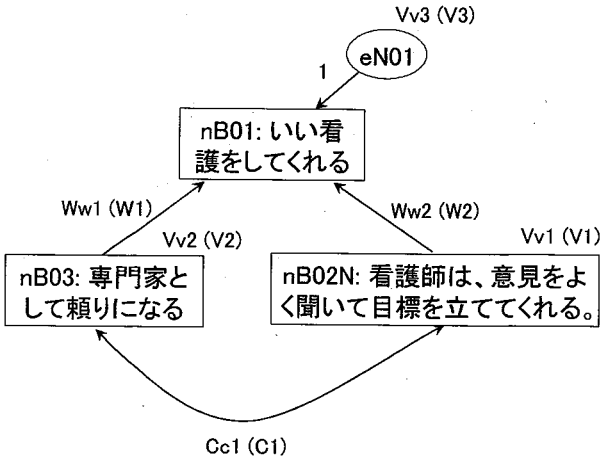

図３．パス図

図３は，それぞれ同位置のパラメーターを対応させて描写したパス図である。（基データ）とＷ１（生成データ），Ｃｃ１（基データ）とＣ１（生成データ）の規則でそれぞれの対応するパラメータにネーミングを行っている。図中では，対応するパラメーターを併記し，"Ｗｗ１（Ｗ１）"と表現している。

表１．パラメーター間の検定統計量

	Ww1	Ww2	Cc1	Vv1	Vv2	Vv3
W1	-0.484	2.439	45.59	25.178	45.58	33.19
W2	-1.907	0.068	25.671	9.069	24.906	14.213
C1	-3.01	-1.772	-0.558	-5.836	-5.241	-5.697
V1	-2.545	-0.996	14.748	0.772	13.026	3.677
V2	-2.83	-1.472	5.395	-3.284	1.865	-2.073
V3	-2.694	-1.244	9.903	-1.338	7.252	0.688

縦方向に基データ、横方向に生成データを配列している

網掛け部の検定統計量の絶対値が、1.942 以上なければ、 5%水準でパラメータ間の差は、有意な差とは言えない

以上の検定結果より，本簡便法は，基データの各質問項目が持つ平均と標準偏差の特性が生成データに反映されていること，および基データの各質問項目間の関係性を反映しているという条件を満たしていると検証できた。

さらに本簡便法は，使用するパソコンおよびソフトウェアーの性能に依存した限界はあるものの，理論的に無限にデータを生成することができ，多量の分析データを生成することが可能である。また本簡便法は，Excel（Microsoft©）を用いて運用できるものであり，利便性の高いものであると考えている。

６．データ生成法のまとめ

開発した基データの特性を再現した簡便なデータ生成法をまとめると，図４に示す通りである。

１）質問項目の内容が質的に他の質問項目の基準となるものを「基準となる変数γ」として選出する。

２）その変数の基データを用いて中心極限定理により，一つのデータを生成する。

３）基準となる変数γを独立変数に，基準とならない変数ｘｉを従属変数として回帰分析を行なう。

４）２）と３）の結果から変数Ｘを補正する。

５）補正された平均ｘと基データの標準偏差を用，中心極限定理により一つのデータｘｉを生成する。

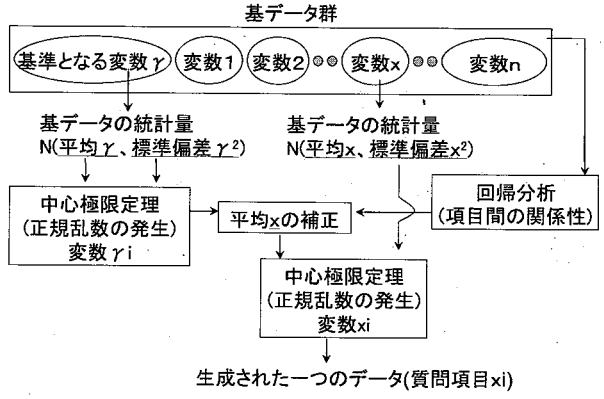

図4. データの生成手順

７．本簡便法の活用方法の検討と活用の留意点

本簡便法は，データサンプリング時の倫理的配慮から，多数のデータが収集できない場合などに活用できる。共分散構造分析のような大規模データを必要とする分析方法を採用することが可能となり，研究協力者の善意によって得られたデータをより有効に活用することができる。また，希少な疾患や事例，試験事業に関する研究など，母集団が小さいテーマの研究，あるいは比較的短い期

間でのデータ収集が必要な場合や，プレテストを基にした予測分析などに応用することができると考えられる。

しかしながら本法は，基準となる変数とある変数の関係性を再現することを繰り返し，データセット内での変数間の関係性を間接的に再現した，あくまでも簡便法であり，変数間の関係性を個々に再現した高度なシミュレーションによるデータ生成に代わるものではない。更に，基データに代わるものではないことは，言うまでもない。

本簡便法の利用に当たっては，これらの原則を踏まえて活用するとともに，基データの収集においては，母集団の特性を正確に反映し，収集したデータに偏りが生じないようにサンプリング方法を十分に吟味して行う必要がある。

8．今後の課題

本簡便法は，基データの特性を正確に反映させるという視点から，今後も洗練化とその妥当性の検討を行って行く必要がある。

また，生成データの結果を安定させるために必要な基データのサンプル数についての検証も必要である。

最後に，この研究にご協力頂いた皆様にあつくお礼申し上げます。

なお，本稿は第25回看護科学学会学術集会において発表したものに，分析を追加したものである。

引用・参考文献

1．有沢　誠，斉藤　鉄也：モデルシミュレーション技法，共立出版，1997

2．井上　正隆：患者の受療者満足の構造分析－看護サービスの提供に対して患者が持つ満足の構造－，病院管理42巻P.93，2005

3．井上　正隆：サービス提供満足の構成要素の分析－看護サービス提供における看護師が持つ満足の構成要素の分析－第9回日本看護管理学会年次大会講演抄録集P.104-105，2005

4．井上　正隆：量的研究における基データのデータ特性を再現したシミュレーションによるデータ生成の簡便法，日本看護科学学会学術集会講演集25号P.189，2005

5．狩野　裕，三浦　麻子：グラフィカル多変量解析—AMOS，EQS，CALISによる目で見る共分散構造分析，現代数学社，2002。

6．薦田　憲久，大川　剛直：システムのモデリングとシミュレーション，コロナ社，1995

7．中西　秀男：シミュレーションの基礎，日新出版，1969

8．奥田　二郎：シミュレーションのＡＢＣ，日本放送出版会，1971

9．坂元　慶行：カテゴリカルデータのモデル分析　共文出版，1985。

10．田部井　明美：SPSS完全活用法—共分散構造分析（Amos）によるアンケート処理，東京図書，2001。

11．山田　覚：医療・看護のためのやさしい統計学　基礎編，東京図書2002。

12．山田　覚：医療・看護のためのやさしい統計学　解析編，東京図書2002。

13．山本　嘉一郎，小野寺　孝義：Amosによる共分散構造分析と解析事例，ナカニシヤ出版，1999。

14．涌井　良幸，涌井　貞美：図解でわかる共分散構造分析，日本実業出版社，2003。