# Let's データ分析コンテストに用いる新擬似ミクロデータの概要

○高橋 行雄[1]，周防 節雄[2]，宮内 亨[3]
[1]BioStat 研究所（株），[2]兵庫県立大学，[3]（独）統計センター

Overview of new pseudo-micro data used in Let's data analysis contest

[1]Yukio Takahashi, [2]Setsuo Suoh, [3]Toru Miyauchi
[1]BioStat Research Co.,Ltd., [2]The Univ. of Hyogo, [2]National Statistics Center

**要旨：** （独）統計センターから提供されていた教育用擬似ミクロデータを用いて SAS ユーザー総会で「Let's データ分析コンテスト」を過去４回開催してきたが，2016 年度末で提供打切りとなった．そこで，2004 年全国消費実態調査の匿名データを用いて，ユーザー会世話人有志が新擬似ミクロデータを新規作成し，2017 年のコンテスト用に供した．新擬似ミクロデータを作成する際には，「匿名データから導いた統計表のみから作成する」ことが（独）統計センターから課せられた必須条件だった．そこで，世帯属性について 14 項目を厳選し，14,246 セルからなる多次元クロス表と集計乗率，収支に関する 203 項目のセル毎の対数平均値と標準偏差，年間収入 3 階級別の主要 21 項目間の対数相関行列を作成して Excel 形式でウェブ上に公開した．この公開情報だけを用いて新擬似ミクロデータを作成した．作成に際しての留意点は以下の通り．①年間収入 3 階級別の 69,131 世帯分の 21 次元正規乱数の作成．②主要 21 項目の 14,246 セル分の対数平均値と標準偏差に対して 21 次元正規乱数を適用し 69,131 世帯分のデータの作成．③主要 21 項目以外の収支 182 項目の 14,246 セルの対数平均値と標準偏差に正規乱数の適用し 69,131 世帯分のデータを作成．④収支金額が 0 円の世帯の割合を保つために一様乱数の使用．⑤下位の収支項目の合計がその上位の項目の収支金額となる様に「足し上げ構造」の保持．⑥正規乱数の適用に際し過剰な発散を防ぐための制約条件の設定．⑦作成された新擬似ミクロデータの収支項目の足し上げ構造の検証．⑧元の匿名データの各種の統計量と比較し，元の性質がどの程度保持されているかを検証．

**キーワード：**全国消費実態調査，擬似ミクロデータ，匿名データ，SAS，JMP，デー分析コンテスト

## 1. はじめに

（独）統計センターから提供されていた教育用擬似ミクロデータを用いて SAS ユーザー総会で「Let's データ分析コンテスト」を過去４回開催してきたが，2016 年度末で提供打切りとなった．これに代わる擬似ミクロデータとして，平成 21 年全国消費実態調査（全消）による一般用ミクロデータが提供されているが，その内容は，教育用擬似ミクロデータの 197 項目に対して，20 項目と 1/10 と大幅に減少し，世帯属性では，例えば，世帯人員属性は「2 人」と「3 人以上」の 2 区分しかなく，利用しずらい．一般用ミクロデータを用いたデータ分析コンテストの実施可能性を検討してみたが，消費支出に関する項目が 10 大費項目に限定されており，公募方式のコンテストに用い

るミクロデータには適さないと判断した.

これまでの教育用擬似ミクロデータを用いた SAS/JMP データ分析コンテストには多数の応募があり，これがきっかけとなって匿名データを用いた実証研究を行う SAS/JMP ユーザーも徐々に増えてきた．統計センターから詳細な一般用ミクロデータの提供が早急に提供される見込みはなく，「Let's データ分析コンテスト」の継続が極めて難しい状況なった.

コンテスト実施の共同オーガナイザである周防が，統計センターの関係者に，匿名データを用いた擬似ミクロデータの作成の可否について問い合わせたところ賛同が得られたので, 2016 年から高橋・周防が使用許可を得ている全消の匿名データの使用目的に，「匿名データを用いた新擬似ミクロデータの作成のための統計表の作成」も追加して申請を行なった.

新擬似ミクロデータ作成は，匿名データ（47,797 世帯分）から導いた複数の統計表のみから作成することと，匿名化の観点から多次元クロス表でのセル度数が 3 以上となることが必須だったので，セル度数 1 及びセル度数 2 に該当する世帯のデータにノイズを入れたデータを加え 69,131 世帯分のデータとし，

① 世帯に関する情報 14 項目と集計乗率

② 14 次元クロス表のセル毎に収支に関する 203 項目の対数変換した平均値と標準偏差

③ 年間収入 3 階級別の主要 21 項目間の相関行列

を作成して，ウェブ上に公開した（高橋・周防）．この情報から新擬似ミクロデータを作成し，SAS データセットと CSV 形式の両方でウェブ上に公開した[1]．詳細手順については、紙幅の関係上、別の機会に詳解する．なお，今回の作業はすべて JMP で行ったが，SAS による作成も近いうちに開始したい.

## 2. 年間収入は対数正規分布に従うのか

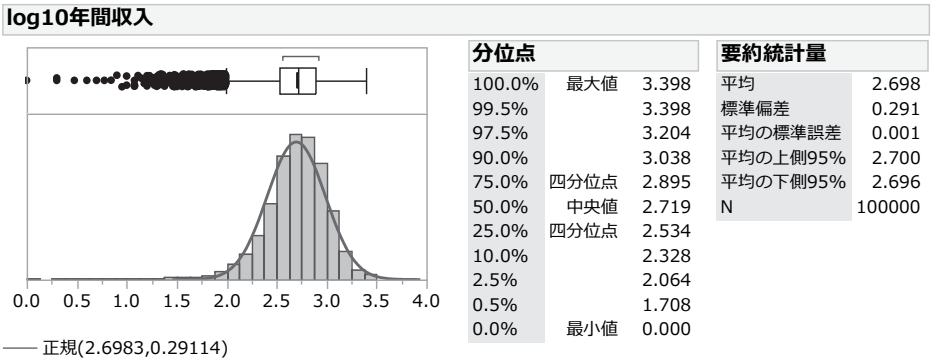
匿名データの集計乗率を用いて，10 万世帯当たりの集計乗率を計算し，それを「度数」として使うと年間収入の平均と標準偏差は（609.7，388.7）となり，変動係数は，63.8%と極めて大きい．その平均と標準偏差の正規分布に従う正規乱数を用いて，47,797 世帯分のデータを擬似的に生成すれば，平均と標準偏差は匿名データとほとんど同じになる．ただし，平均と標準偏差の関係から，実際にはありえない「年間収入がマイナス」の世帯が多数発生してしまう.

年間収入は，高いほう（右に裾を引く）対数正規分布に従うと仮定できるのか．図 1a は，匿名データに対して正規分布をあてはめた結果で，左側に大きく裾を引いていることが観察される．図 1b は，正規乱数データに対する正規分布のあてはめでヒストグラム上にきれいに乗っている.

図 1a の匿名データは，右側のひげを超える点が全くない．これは，2500 万円を超える年間収入は 2500 万円とするトップコーディングが 268 世帯（0.56%）に施されているためであり，99.5%分位点と最大値が 3.398（2500 万円）と同じであることからも確認できる．図 1a の要約統計量の（平均=2.698, 標準偏差=0.291）を用いた正規乱数で生成したデータでは，図 1b に示すようにきれいに左右対称となり，99.5%点は，3.469（2944 万円）と若干大きくなる．2.5%点では log10（年間収入）で 2.064（116 万円），乱数データで 2.134（134 万円）と食い違いが生じている.

---

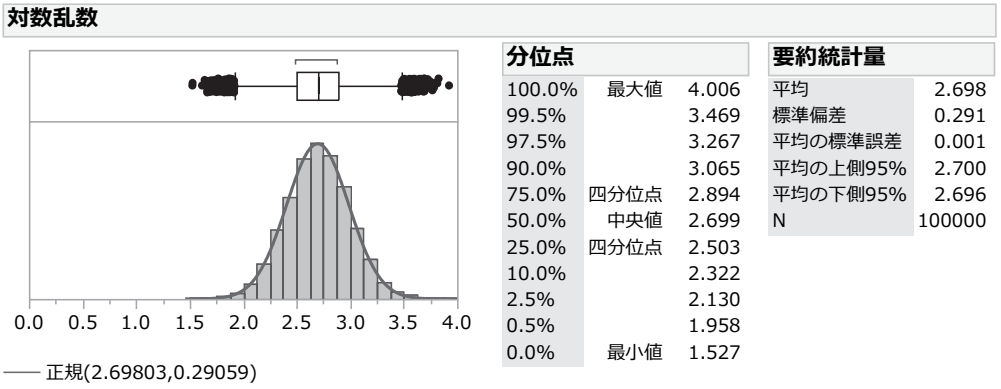
[1] SAS データセット ⇒ http://mighty.gk.u-hyogo.ac.jp/confidential/Zensho2004GijiMicroData.zip
CSV 形式 ⇒ http://mighty.gk.u-hyogo.ac.jp/confidential/Zensho2004GijiMicroDataCSV.zip




| 分位点 | | |
|---|---|---|
| 100.0% | 最大値 | 3.398 |
| 99.5% | | 3.398 |
| 97.5% | | 3.204 |
| 90.0% | | 3.038 |
| 75.0% | 四分位点 | 2.895 |
| 50.0% | 中央値 | 2.719 |
| 25.0% | 四分位点 | 2.534 |
| 10.0% | | 2.328 |
| 2.5% | | 2.064 |
| 0.5% | | 1.708 |
| 0.0% | 最小値 | 0.000 |

| 要約統計量 | |
|---|---|
| 平均 | 2.698 |
| 標準偏差 | 0.291 |
| 平均の標準誤差 | 0.001 |
| 平均の上側95% | 2.700 |
| 平均の下側95% | 2.696 |
| N | 100000 |

図 1a　常用対数変換した年間収入の分布の特徴




| 分位点 | | |
|---|---|---|
| 100.0% | 最大値 | 4.006 |
| 99.5% | | 3.469 |
| 97.5% | | 3.267 |
| 90.0% | | 3.065 |
| 75.0% | 四分位点 | 2.894 |
| 50.0% | 中央値 | 2.699 |
| 25.0% | 四分位点 | 2.503 |
| 10.0% | | 2.322 |
| 2.5% | | 2.130 |
| 0.5% | | 1.958 |
| 0.0% | 最小値 | 1.527 |

| 要約統計量 | |
|---|---|
| 平均 | 2.698 |
| 標準偏差 | 0.291 |
| 平均の標準誤差 | 0.001 |
| 平均の上側95% | 2.700 |
| 平均の下側95% | 2.696 |
| N | 100000 |

図 1b　正規乱数データ（平均=2.698，標準偏差=0.291）の分布

データ分析コンテストでは，年間収入について各種の統計解析が行なわれることも想定されるので，このようなわずかな分布のゆがみも再現することが望ましいと考える．図 1a ヒストグラムから年間収入の分布は，視覚的にほぼ対数正規分布に従っていると見なせるかもしれないが，図 2 に示す正規分位点プロットを見れば，年間収入が低くなるにつれて，斜めの直線からの乖離が大きくなり，対数正規分布に従っているとみなすことは躊躇せざるを得ない．

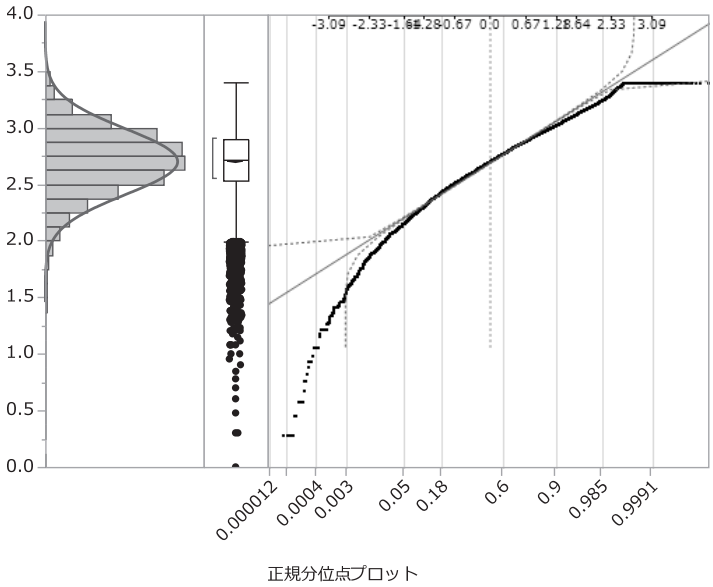



図 2 年間収入（対数）についての正規確率プロット

このような場合には，複数の母集団が内在していると仮定し，そこからランダムに 47,797 世帯がサンプリングされたとし，各母集団の(混合割合，平均，標準偏差)を最尤法によって推定することが可能である．

母集団が 2 つあると仮定した場合と，3 つあると仮定した場合の統計量を図 3 に示す．母集団が 2 つと仮定した場合に，「2 重正規分布のあてはめ」欄の割合 $\pi 1$ が 13.4%の小集団で，平均値は位置 $\mu 1$ =2.388(244 万円)であり，割合 $\pi 2$ が 86.6%を占める大集団で，平均値は 2.746(557 万円)である．ヒストグラム上の確率密度関数のあてはめも母集団が 1 つと仮定した場合に比べ改善している．さらに母集団が 3 つと仮定した場合には，「3 重正規分布のあてはめ」欄の割合(1.7%，8.0%，90.3%)の集団が分離されている．

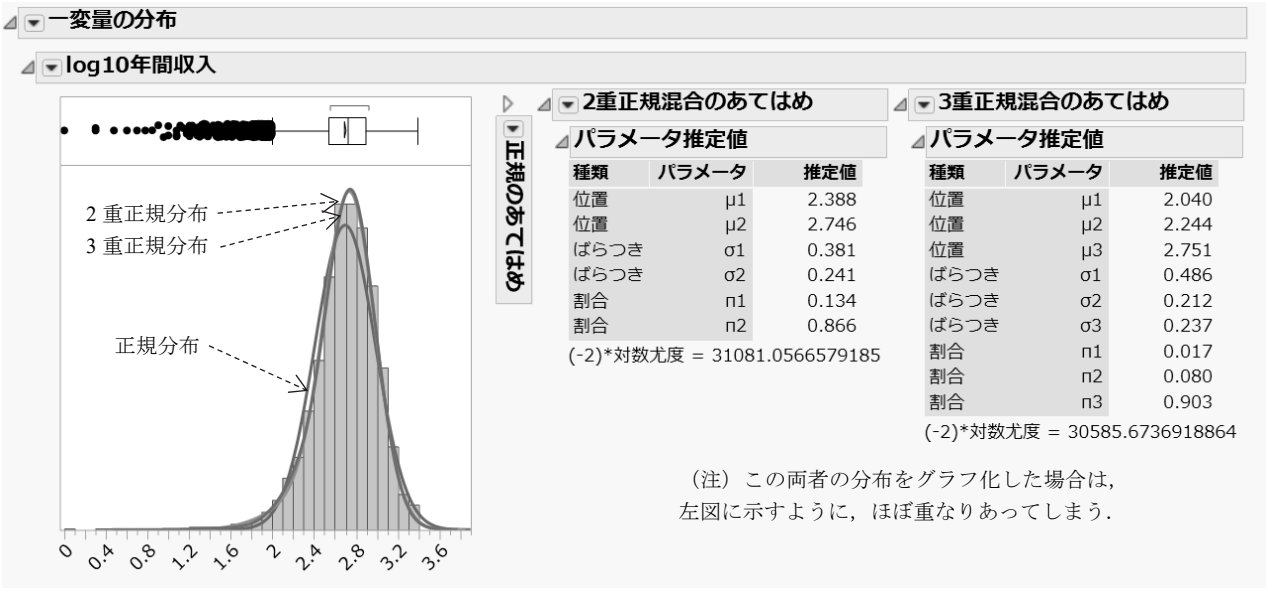



図 3 混合正規分布のあてはめ

統計的にどちらのあてはめが望ましいのかは，マイナス 2 倍の対数尤度を用いて評価する．表 1 に示すように，母集団=1 の場合は，(-2)*対数尤度=36995.97 で，母集団=2 の場合は，(-2)*対数尤度=31081.06 となり，5914.91 の減少となる．母集団=3 の場合は 30585.67 で，母集団=2 の場合に比べ 495.39 小さくなる，母集団=4 の場合は，母集団=3 の場合に比べ 45.82 小さくなる．母集団=5 の

表 1　仮定する母集団の数による対数尤度の差および最小母集団の統計量

| | | | 最小の母集団 | | |
|---|---|---|---|---|---|
| 母集団数 | (-2)対数尤度 | 対数尤度の差 | 構成比 | 対数母平均 | 母平均(万) |
| 1 | 36995.97 | - | 100.0% | 2.698 | 498.9 |
| 2 | 31081.06 | -5914.91 | 13.4% | 2.388 | 244.3 |
| 3 | 30585.67 | -495.39 | 1.7% | 2.040 | 109.6 |
| 4 | 30539.85 | -45.82 | 0.021% | 0.513 | 3.3 |
| 5 | 30539.84 | -0.01 | 0.025% | 0.587 | 3.9 |
| | $\chi^2(df=2, 5\%)$ | 5.99 | | | |

場合には変化が 0.01 である．統計的には，$\chi^2$=5.99(自由度 2, 上側 5%点)以下ならば，母集団数の増加に意味がないので，母集団=4 が統計的に支持される．しかしながら，母集団=4 の場合，最小母集団の構成比は，0.021%と極めて小さい集団なので，母集団=3 とするのが現実的と思われる．その場合，年間収入の母平均は，それぞれ 110 万円，175 万円，563 万円となる．

年間収入以外の 202 項目についてもヒストグラムの視覚的な検討結果から，単一の対数正規分布とはみなすことができないと判断した．それ故に，世帯属性毎に，14 次元クロス表の 14,246 セルに対して 203 項目の(データ数，対数平均，標準偏差)の統計表を作成・公表した．

次節以降で，公表した統計表から新擬似ミクロデータを作成した過程を示す．

## 3. 正規乱数を用いた復元

公表した(14,246×3 レコード×収支 203 項目)の統計表から，(69,131 レコード×収支 203 項目のデータの作成が最初のステップである．表 2 に，統計表の一部を示す．

表 2　収支項目に関する統計表（抜粋）

| 14次元番号 | ラベル | Y001 | Y002 | Y003 | Y004 | Y005 | (中略) | Y038 | Y039 | Y040 | Y041 | (中略) | Y203 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Z01 | Z02 | Z03 | | | | Z06 | Z07 | Z08 | Z09 | | Z21 |
| | | 年間収入 | 収入総額 | 実収入 | 経常収入 | 勤め先収入 | (収入項目) | 支出総額 | 実支出 | 消費支出 | 食料 | (支出項目) | 繰越金 |
| 1 | 件数* | 3 | 3 | 3 | 3 | 3 | | 3 | 3 | 3 | 3 | | 3 |
| 1 | log10平均 | 2.192 | 5.226 | 5.112 | 5.119 | 5.104 | | 5.224 | 4.764 | 4.756 | 3.910 | | 4.719 |
| 1 | log10_SD | 0.005 | 0.015 | 0.010 | 0.034 | 0.009 | | 0.023 | 0.011 | 0.016 | 0.008 | | 0.023 |
| 2 | 件数 | 3 | 3 | 3 | 3 | 3 | | 3 | 3 | 3 | 3 | | 3 |
| 2 | log10平均 | 2.430 | 5.615 | 5.338 | 5.341 | 5.325 | | 5.627 | 5.303 | 5.207 | 4.557 | | 4.014 |
| 2 | log10_SD | 0.006 | 0.009 | 0.010 | 0.012 | 0.013 | | 0.010 | 0.010 | 0.014 | 0.011 | | 0.019 |
| | (中略) | | | | | | | | | | | | |
| 14246 | 件数 | 3 | 3 | 3 | 3 | 3 | | 3 | 3 | 3 | 3 | | 3 |
| 14246 | log10平均 | 3.112 | 6.16 | 5.921 | 5.912 | 5.771 | | 6.162 | 5.751 | 5.635 | 5.076 | | 5.052 |
| 14246 | log10_SD | 0.011 | 0.007 | 0.014 | 0.013 | 0.002 | | 0.011 | 0.013 | 0.027 | 0.004 | | 0.005 |
| | | * 件数＝それぞれの収支項目の金額が0円であるケースを除いた出現頻度である. | | | | | | | | | | | |

① 14 次元表の 14,246 レコードのセル度数を用いて，該当するレコードをセル度数分複製し，順次縦方向に連結する．最終的には，セル度数の合計 69,131 レコードを持つファイルを作成する．

② 公表した統計表を転置して(14,246×203)×3 のファイルを作成する．

③ ①と②のファイルに対し，セル番号でマッチマージをして②の情報を①に付与し，(69,131×203)×3 のファイルを作成する．

④ ③のファイルに正規乱数を追加して，この正規乱数を使って「対数平均＋標準偏差×正規乱数」を計算する．ただし，標準偏差は±2 の範囲内に制限し，これに 0.5 を掛けて変動が±1 となるように制限を加えた．

⑤ ④のファイルを世帯毎に転置し，69,131×203 の擬似データのファイルを作成する．

⑥ 一様乱数を用いて欠測値処理をする．（これについては，第 6 節を参照）

## 4. 互いに相関を持つ主要 21 項目

主要 21 項目について，年間収入 3 階級別に 21×21 の相関係数行列を公表している．互いに相関を持つ正規乱数の作成は，相関係数行列をコレスキー分解した行列と正規乱数行列の積で求められ

る．手順を可視化するために JMP のスクリプトを用いて例示する．

まず，3 項目の 3×3 の相関係数行列を roh とする．Cholesky 関数でコレスキー分解を行なうと下三角行列 chol が得られる．

**roh（相関行列）**

| | | |
|---|---|---|
| 1 | 0.8 | 0.5 |
| 0.8 | 1 | 0.7 |
| 0.5 | 0.7 | 1 |

➔ `chol=Cholesky(roh);` ➔

**chol（コレスキー分解）**

| | | |
|---|---|---|
| 1 | 0 | 0 |
| 0.8 | 0.6 | 0 |
| 0.5 | 0.5 | 0.707 |

次に 3×5 の正規乱数行列 ysnorm を Random Normal 関数で生成する．

`ysnorm=J(3,5,Random Normal());` ➔

**ysnorm（正規乱数・相関 0）**

| | | | | |
|---|---|---|---|---|
| 0.274 | -0.224 | 0.219 | -0.420 | 0.246 |
| 0.908 | 1.485 | 1.663 | -1.682 | -0.761 |
| 0.629 | 1.432 | 0.360 | -0.023 | 0.467 |

下三角行列 chol と正規乱数行列 ysnorm の積を求め，互いに相関を持つ行列を計算し，転置（JMP のスクリプトでは「`」が転置記号）して．5×3 の 3 次元正規乱数行列 ymnorm を求める．

`ymnorm=(chol*ysnorm)`;` ➔

**ymnorm**

| | | |
|---|---|---|
| **0.274** | **0.764** | **1.036** |
| -0.224 | 0.712 | 1.643 |
| 0.219 | 1.173 | 1.196 |
| -0.420 | -1.345 | -1.067 |
| 0.246 | -0.260 | 0.073 |

最初に， chol の 1 行目 [1.0　0　0] と ysnorm の 1 列目 [0.274　0.908　0.629]$^T$ の積和が ysnorm の 1 行 1 列目に 0.272 と計算され，元の正規乱数のままとなる．次に，chol の 2 行目と ysnorm の 1 列目の積和なので，

$$0.8\times0.274+0.6\times0.908=0.764$$

ysnorm の 1 行 2 列目に 0.764 が得られる．さらに，chol の 3 行目 [0.5　0.5　0.707] に ysnorm の 1 列目 [0.274　0.908　0.629]$^T$ の積和によって，1.036 が得られる．

| （相関を加味した正規乱数） | |
|---|---|
| **0.274** | **=1　x0.274 + 0　x0.908 + 0　x0.629** |
| **0.764** | **=0.8x0.274 + 0.6x0.908 + 0　x0.629** |
| **1.036** | **=0.5x0.274 + 0.5x0.908 + 0.0707x0.629** |

この操作を，ysnorm の 2 行目から 5 行目まで繰り返し，5×3 の 3 次元正規乱数行列 ymnorm が計算されている．実際に 21 次元の相関係数行列を 69,131 世帯分作る JMP スクリプトを次ページに示す．

このスクリプトを年齢 3 階級毎に，公表した相関係数行列をスクリプトに挿入し，69,131 行 21 列の互いに相関を持つ正規乱数行列を計算し，この行列をファイルに出力する．表 3 に第 1 階級のファイルの一部を示す．

```
//  相関を持つ正規乱数
nrows=69131 ;
seed=1234561 ;
Random Reset(seed);
roh=
[   1.000  0.462  ... 0.133,
    0.462  1.000  ... 0.217,
      :      :         :
    0.133  0.217  ...1.000]
;
ysnorm=J(Ncol(roh),Nrows, Random Normal());
chol=Cholesky(roh);
ymnorm=(chol*ysnorm)`;
dt = New TAble("ran") << Set Matrix(ymnorm);
```

表 3　下位 1/3 階級用の 69,131 行 21 列の互いに相関を持つ正規乱数

GIJI_年収区分_1低_21ran - JMP

ファイル(F)　編集(E)　テーブル(T)　行(R)　列(C)　実験計画（DOE)(D)　分析(A)　グラフ(G)　ツール(O)　表示(V)　ウィンドウ(W)　ヘルプ(H)

GIJI_年収区分…

列(21/0)

行

すべての行　69,131

選択されている行　0

| | 列1 | 列2 | 列3 | 列4 | 列5 | 列6 | 列7 | 列8 | 列9 | 列10 | 列11 | 列12 | 列13 | 列14 | 列15 | 列16 | 列17 | 列18 | 列19 | 列20 | 列21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.23 | 0.26 | 1.29 | -0.02 | -0.46 | 0.04 | 1.82 | 1.82 | -0.53 | 1.67 | -0.93 | -0.35 | 0.31 | 0.02 | 1.23 | 1.19 | 0.77 | 1.14 | 1.36 | -0.46 | -1.02 |
| 2 | -0.29 | 0.47 | 1.07 | 0.92 | -1.17 | 0.73 | 1.08 | 1.06 | 0.13 | 1.08 | 0.08 | 1.16 | 2.06 | 0.88 | 1.76 | 0.03 | 0.22 | 1.01 | -0.23 | 0.16 | -1.04 |
| 3 | -0.37 | 0.67 | 0.33 | 1.38 | 1.84 | 1.04 | 0.21 | 0.29 | 0.30 | -0.69 | -1.40 | 0.29 | -1.42 | 1.52 | 0.10 | -1.52 | 0.21 | 1.69 | -0.99 | -0.04 | 1.92 |
| 4 | -1.22 | -0.28 | -0.16 | -0.99 | -0.46 | -0.63 | 0.58 | 0.83 | 0.16 | -0.98 | -0.32 | -0.10 | 0.69 | -0.08 | 0.23 | 0.07 | -0.29 | 2.62 | -1.86 | -1.67 | 0.75 |
| 5 | 0.60 | 0.90 | 0.27 | 0.52 | 0.16 | 1.01 | 1.41 | 1.15 | 1.11 | 0.20 | -0.14 | 1.18 | 2.39 | 0.62 | 0.60 | -0.54 | 1.60 | 1.52 | 1.32 | 0.54 | -0.62 |

主要 21 項目に対し多次元正規乱数による誤差変動の与え方を以下に示す.

① 年間収入 3 階級の(69,131 世帯)×21 項目の多変量正規乱数ファイル毎に世帯毎に転置し(69,131×21)×1 とする.

② 年間収入 3 階級毎に転置されたデータを列方向に併合し,(69,131×21)×3 の乱数ファイルとする.

③ 世帯毎の 3 階級の符号に対応した乱数列を用いる.

④ 主要 21 項目は，203 項目の一部なので 203 項目のファイルに上書きする.

## 5. 住居費および教育費など支出金額が 0 円となる世帯が多数ある場合

匿名データ 47,797 世帯を，集計乗率を考慮して 10 万世帯とした場合に，住居費が支出されている世帯は，66,113 世帯あり，支出金額が 0 円の世帯が 1/3 程度含まれている．図 4a に示すように，分布は左側に大きく裾を引いていて 3 重混合対数正規分布のあてはめが支持される．母集団の月当たり住居費の母平均は,「3 重正規混合のあてはめ」欄の $\mu 1$ , $\mu 2$ , $\mu 3$ を用いて, それぞれ $10^{2.547}$=352, $10^{4.113}$=12,972，$10^{4.736}$ =54,450 円/月と推定される.

図 4b に示すように，教育費が支出されている世帯は，10 万世帯中 28,540 世帯あり，欠測値が 3/4 程度含まれている．分布は左右ともに裾を引いていて，3 重混合対数正規分布のあてはめが支持さ

れる．母集団の平均値は，住居費と同様に，それぞれ $10^{3.887}$=7,709，$10^{4.440}$=27,542，$10^{5.078}$=119,674 円/月と推定される．


図 4a　匿名データ 47,797 世帯の 10 万世帯での対数住居費の分布


図 4b　匿名データ 47,797 世帯の 10 万世帯での対数教育費の分布

住居費には，家賃地代や修繕費などが含まれているが，持家でその年に修繕費が発生しなければ支出金額は 0 円となり，教育費は，対象になる家族がいなければ 0 円である．公表した統計量の算出に際して，常用対数変換を施したので 0 円は欠測値となり，(69,131 世帯×203 項目)×3 統計量のファイルに付加された世帯毎・項目毎の欠測値を含むデータ数を $N_i$ ， $i$=1,2,...,14246 セルとし，欠測値以外のデータ数を $n_{ij}$ ， $j$=1,2,..., $N_i$ ， $n_{i1}$= $n_{i2}$=⋯= $n_{iNi}$ として，欠測値の割合を保つ擬似データを，以下の基準で生成する．

① データ数が $n_{ij}$=0 の場合，対数平均値は欠測値なので，$N_i$ 個の擬似データ全てを欠測値．
② データ数が $n_{ij}$=1 の場合には，$N_i$ 個中，1 個は擬似データがあり，$N_i$ - 1 個は欠測値．
③ 0～1 の一様乱数を $u_{ij}$ とし，$u_{ij}$ が $n_{ij}$ /$N_i$ の比より小さければ擬似データがあり，大きければ欠測値．
④ 表 3 に示すように $N_i$ が 5 で，データ数 $n_{ij}$ が 3 の場合，対数平均と標準偏差は，5 世帯が全て同じなので 5 世帯中 2 世帯は欠測値．
⑤ 一様乱数 $u_{ij}$ を発生させて，3/5=0.60 以下ならば擬似データがあり，0.60 以上ならば欠測値(表 4)．

表 4.　一様乱数による教育費が 0 円の世帯の割合の確保

| 世帯番号 | 14次元番号 | レコード数 $N_i$ | 繰返し $j$ | 項目名 | データ数 $n_{ij}$ | 対数平均 | 対数SD | 一様乱数 | $n_{ij}/N_i$ | 判定 | 扱い |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 9999 | 8888 | 5 | 1 | 教育費 | 3 | 4.00 | 0.10 | 0.70 | 0.60 | × | 欠測値 |
| 9999 | 8888 | 5 | 2 | 教育費 | 3 | 4.00 | 0.10 | 0.20 | 0.60 | ○ | 採用 |
| 9999 | 8888 | 5 | 3 | 教育費 | 3 | 4.00 | 0.10 | 0.80 | 0.60 | × | 欠測値 |
| 9999 | 8888 | 5 | 4 | 教育費 | 3 | 4.00 | 0.10 | 0.39 | 0.60 | ○ | 採用 |
| 9999 | 8888 | 5 | 5 | 教育費 | 3 | 4.00 | 0.10 | 0.50 | 0.60 | ○ | 採用 |

実際に作成された新擬似ミクロデータ 69,131 世帯の住居費と教育費について，10 万世帯での分布をそれぞれ図 5a と図 5b に示す．住居費が支出されている世帯数は，図 4a と比較して，66,113 世帯から 61,832 世帯になり，新擬似ミクロデータで 3 重正規混合分布の形状も保持されている．


図 5a　新擬似ミクロデータの 10 万世帯での対数住居費

教育費が支出されている世帯数は，図 4b と比較すると 28,540 世帯から 29,066 世帯になり，ほぼ 0 円世帯の割合は保たれている．また，分布も 3 重正規混合分布の形状がやや滑らかになるが，(-2) 対数尤度の大きな減少もあり，新擬似ミクロデータで保持されている．


図 5b　新擬似ミクロデータの 10 万世帯での対数教育費

## 6. 足し上げ構造

表 4 に示す「消費支出」は，その下の項目「食料，住居，...，その他の消費支出」の 10 大消費項目の金額の合計が一致するように統制されており，これを「足し上げ構造」と呼ぶことにする．

この様な階層構造が他にも多くあり，表 5 に，主要 21 項目間の足し上げ構造について，「区分 $i$」と「足し上 $i$」を用いて表す．例えば「区分 2」では，21 と 22 の 2 種類の足し上げ構造があり，それぞれ，「足し上 2」の(0, 1)で親子関係が示されている．同じ区分内の親(0)は，子(1)の合計金額と一致する．「区分 3」の 33 では，「実支出」を親とし，子が「消費支出」と「非消費支出」の 2 変数から成る．この表に示した表記法は，新擬似ミクロデータに含まれる足し上げ構造を考慮して計算する際に使えるにように，今回考案した．

表 5　主要 21 項目足し上げ構造

| | | | レベル2 | | レベル3 | | レベル4 | | 世帯1 | | | 世帯2 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Z | V | 区分2 | 足上2 | 区分3 | 足上3 | 区分4 | 足上4 | データ | 足し上げ | | データ | 足し上げ | |
| | | V0003_調査年 | | | | | | | 2004 | | | 2004 | | |
| 階層 | | V0004_No | | | | | | | 1 | | | 2 | | |
| 2 | Z01 | V0399_年間収入 | | | | | | | 76 | | | 738 | | |
| 2 | Z02 | V0400_収入総額 | 21 | 0 | | | | | 353,255 | 353,255 | o | 1,282,209 | 1,282,209 | o |
| 3 | Z03 | V0401_実収入 | 21 | 1 | | | | | 309,924 | | | 354,483 | | |
| 3 | Z04 | V0439_実収入以外の収入 | 21 | 1 | | | | | 3,715 | | | 672,663 | | |
| 3 | Z05 | V0452_繰入金 | 21 | 1 | | | | | 39,615 | | | 255,062 | | |
| 2 | Z06 | V0453_支出総額 | 22 | 0 | | | | | 353,255 | 353,255 | o | 1,282,209 | 1,282,209 | o |
| 3 | Z07 | V0454_実支出 | 22 | 1 | 33 | 0 | | | 256,381 | 256,381 | o | 485,670 | 485,670 | o |
| 4 | Z08 | V0455_消費支出 | | | 33 | 1 | 45 | 0 | 256,381 | 256,381 | o | 470,916 | 470,916 | o |
| 5 | Z09 | V0456_食料 | | | | | 45 | 1 | 68,774 | | | 90,732 | | |
| 5 | Z10 | V0498_住居 | | | | | 45 | 1 | 52,138 | | | 16,840 | | |
| 5 | Z11 | V0504_光熱・水道 | | | | | 45 | 1 | 17,125 | | | 22,710 | | |
| 5 | Z12 | V0509_家具・家事用品 | | | | | 45 | 1 | 11,387 | | | 12,172 | | |
| 5 | Z13 | V0519_被服及び履物 | | | | | 45 | 1 | 12,265 | | | 3,208 | | |
| 5 | Z14 | V0537_保健医療 | | | | | 45 | 1 | 2,132 | | | 17,123 | | |
| 5 | Z15 | V0542_交通・通信 | | | | | 45 | 1 | 43,016 | | | 12,012 | | |
| 5 | Z16 | V0553_教育 | | | | | 45 | 1 | 0 | | | 0 | | |
| 5 | Z17 | V0557_教養娯楽 | | | | | 45 | 1 | 12,614 | | | 16,505 | | |
| 5 | Z18 | V0567_その他の消費支出 | | | | | 45 | 1 | 36,930 | | | 279,615 | | |
| 4 | Z19 | V0598_非消費支出 | | | 33 | 1 | | | 0 | | | 14,754 | | |
| 3 | Z20 | V0609_実支出以外の支出 | 22 | 1 | | | | | 32,848 | | | 485,652 | | |
| 3 | Z21 | V0622_繰越金 | 22 | 1 | | | | | 64,026 | | | 310,887 | | |

表 6 に，消費支出の細目の足し上げ構造を示す．「穀類」は，「米，パン，めん類，他の穀類」の合計となっていることが確認できる．

足し上げ構造は，下位の区分から上位の区分に対して定義されているが，正規乱数を用いて変動を与えているので，足し上げ構造に若干の揺れが生じている．下位から上位への足し上げは，誤差変動の伝播により上位の項目の変動が大きくなりすぎる傾向がある．そこで，上位を先に固定してから下位方向に調整する「逆」足し上げ計算を行う．

子の金額を $z_{i1j}$（$i$ は各レベルの区分番号，1 は足し上げの子のコード，$j$ は項目番号），親の金額を $z_{i0}$（$i$ は各レベルの区分番号，0 は足し上げの親のコード），$N$ 個の子の項目の金額の合計を $z_{i1\cdot}$ とした時に，「逆足し上げ」により調整された子(項目番号 $j$)の金額 $z'_{i1j}$ を，

$$z'_{i1j} = \left( \frac{z_{i1j}}{z_{i1\cdot}} \right) z_{i0}$$

で定義する．こうすれば，

$$z_{i0} = \sum_{j=1}^{N} z'_{i1j}$$

表 6　食料における足し上げ構造（抜粋）

| | | | | レベル2 | | レベル3 | | レベル4 | | レベル5 | | レベル6 | | 世帯1 | | | 世帯2 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 階層 | 新擬似Y名 | 匿名V名 | 区分 | 足上 | 区分 | 足上 | 区分 | 足上 | 区分 | 足上 | 区分 | 足上 | データ | 足し上げ | | データ | 足し上げ | |
| 39 | 3 | Y039 | V0454_実支出 | 22 | 1 | 33 | 0 | | | | | | | 256,381 | 256,381 | o | 485,670 | | |
| 40 | 4 | Y040 | V0455_消費支出 | | | 33 | 1 | 45 | 0 | | | | | 256,381 | 256,381 | o | 470,916 | 470,916 | o |
| 41 | 5 | Y041 | V0456_食料 | | | | | 45 | 1 | 503 | 0 | | | 68,774 | 68,774 | o | 90,732 | 90,732 | o |
| 42 | 6 | Y042 | V0457_穀類 | | | | | | | 503 | 1 | 602 | 0 | 11,957 | 11,957 | o | 10,362 | 10,362 | o |
| 43 | 7 | Y043 | V0458_米 | | | | | | | | | 602 | 1 | 6,359 | | | 4,627 | | |
| 44 | 7 | Y044 | V0459_パン | | | | | | | | | 602 | 1 | 3,719 | | | 3,209 | | |
| 45 | 7 | Y045 | V0460_めん類 | | | | | | | | | 602 | 1 | 1,712 | | | 1,962 | | |
| 46 | 7 | Y046 | V0461_他の穀類 | | | | | | | | | 602 | 1 | 168 | | | 565 | | |
| 47 | 6 | Y047 | V0462_魚介類 | | | | | | | 503 | 1 | 603 | 0 | 1,197 | 1,197 | o | 16,122 | 16,122 | o |
| 48 | 7 | Y048 | V0463_生鮮魚介 | | | | | | | | | 603 | 1 | 1,197 | | | 10,718 | | |
| 49 | 7 | Y049 | V0464_塩干魚介 | | | | | | | | | 603 | 1 | 0 | | | 2,504 | | |
| 50 | 7 | Y050 | V0465_魚肉練製品 | | | | | | | | | 603 | 1 | 0 | | | 1,142 | | |
| 51 | 7 | Y051 | V0466_他の魚介加工品 | | | | | | | | | 603 | 1 | 0 | | | 1,758 | | |

となり，子の項目の合計が，親の金額と一致する．実際の計算手順は以下の通り．

① 69,131 世帯毎に 203 項目を転置したファイルを作成する．
② 表 6 に示した階層データ構造を表現している「レベル 2 の区分と足し上, . . . , レベル 6 の区分と足し上」を変数名で①とマッチマージし（69,131 世帯×203 項目）レコード×「収支金額＋5 レベル×2（区分・足し上)」のファイルを作成する．
③ ②のファイルに対して「世帯×区分 $i$×足し上げ (0,1)」別に金額の合計を計算し，世帯毎に $z_{i0}$ と $z_{i1.}$ .を得る．
④ ②と③のファイルを，世帯番号，区分 $i$ でマッチマージすることによって，①のファイルに $z_{i1.}$ を付加して， $z'_{i1j}$ を計算する．
⑤ すべての区分に対して手順③を繰り返す.

なお，一様乱数を用いて 0 円データを修正すると，足し上げ構造の親子関係に不具合が発生するケースがある．すなわち，親が欠測値なのに，子に値がある，逆に親に値があるが，子はすべて欠測値の場合が生ずる．このような場合には，逆足し上げ計算はできないので，親・子の一方が欠測値の場合には，事後処理として，両方とも 0 円に置き換える．

## 7. 新擬似ミクロデータと匿名データとの照合および足し上げ構造の検証

新擬似ミクロデータと匿名データとの照合するために，世帯属性の 14 項目の符号について構成比を算出し比較した結果，±5%程前後の相違があった．収支の 203 項目の対数平均の総平均は，匿名データで 3.811，新擬似ミクロデータで 3.765 と 1.2%小さかった．元の金額にすると 10.0%の差となる．元の金額で 50%を超える差異が 5 項目にあったが，ほぼ匿名データの特質は再現できたと思われる．

新擬似ミクロデータの 203 項目の金額について，すべて四捨五入により整数化を行ったので，足し上げ構造に若干のくるいが生じている．全足し上げ 3,387,419 箇所中，最大で±4 円は 13 箇所，±1 円差が 663,196 箇所 19.6%あった．今回は，急いで新擬似ミクロデータを提供しなければならない状況であったので，「足し上げには，四捨五入による不整合がある」との注を入れることに留めて，

以下の URL に新擬似ミクロデータを公開した（高橋・周防・宮内）．将来的には，今の数字を使って，再度，通常の足し上げ計算をして，親の値を書き換えることも検討している．

SAS データセット　http://mighty.gk.u-hyogo.ac.jp/confidential/Zensho2004GijiMicroData.zip
CSV 形式　　http://mighty.gk.u-hyogo.ac.jp/confidential/Zensho2004GijiMicroDataCSV.zip

## 8. 今後の展望

新擬似ミクロデータは，誰でも何時でも自由にダウンロードして使えるが，約 7 万件のデータを用いた統計解析は，誰にでも容易にできるという訳ではない．どの様な解析がいかにしたらできるのかは，無料でダウンロードできる web 上の SAS ユーザー総会の論文集に収録されている「Let's データ分析コンテスト」のこれまでの優秀賞の論文が参考になる．今後，コンテストにチャレンジするユーザーが増えることを願っている．

今年の第 5 回目コンテストの実施に間に合うように，新擬似ミクロデータを提供できたことに安堵している．ただ，急造したために，一部にケアレスミスがあったことは，深くお詫びしたい．今回の 2004 年データに加え，1989 年，1994 年，1999 年の全国消費実態調査についても，匿名データを用いて擬似ミクロデータを作成し，来年以降の「Let's データ分析コンテスト」に供したい．


## 謝　辞

全国消費実態調査の匿名データの利用のために，必要な予算上の措置を賜った SAS Institute Japan（株），ならびに，匿名データの利用に便宜を図って頂いた（独）統計センターのそれぞれの関係各位に深く感謝申し上げます．


## 参 考 文 献


1） 周防節雄（2015）全国消費実態調査の匿名データとその符号表から自動的に SAS のデータセット，変数ラベルと変数フォーマットを作成する SAS プログラム，『SAS ユーザー総会 2015 論文集』，pp257-278.
2） 高橋行雄（2015）統計センターの匿名データ 13 万件を用いた統計解析の実践，『SAS ユーザー総会 2015 論文集』，pp145-164.
3） 高橋行雄（2016）JMP による第4回 Let's データ分析の規定課題の解析，『SAS ユーザー総会 2015 論文集』，pp313-329.
4） 周防節雄（2016）SAS ユーザー総会 2016 における「Let's データ分析第4回ミクロデータ分析コンテスト」の規定課題の SAS プログラム解説，『SAS ユーザー総会 2016 論文集』，pp330‐347.
5） 高橋行雄（2016）統計センター提供の教育用擬似ミクロデータを用いた SAS/JMP によるデータ分析コンテスト，https://www.nstac.go.jp/services/pdf/161125_3-3.pdf.