社会システム分析のための統合化プログラム２１

― 乱数生成と検定 ―

福井正康・*孟紅燕・*呉夢・*崔永杰

福山平成大学経営学部経営学科
*福山平成大学大学院経営学研究科経営情報学専攻

概要

我々は教育分野での利用を目的に社会システム分析に用いられる様々な手法を統合化したプログラム College Analysis を作成してきた。今回は、様々なシミュレーションや統計的な母数推定などに用いられる乱数生成とその検定についてプログラムを作成した。乱数生成では、通常の乱数個別の方法に加え、マルコフ連鎖モンテカルロ法を用いた汎用的に利用できるプログラムについても紹介する。また乱数の検定には、ヒストグラム、p-p プロット、適合度検定、コルモゴロフースミルノフ検定が含まれている。

キーワード

College Analysis，社会システム分析，統計，乱数，マルコフ連鎖モンテカルロ法，MCMC

URL: http://www.heisei-u.ac.jp/ba/fukui/

１．はじめに

これまで、我々は社会システム分析ソフトウェア College Analysis において、統計分析、数学、経営科学、意思決定手法などを中心にプログラムを作成してきたが、今回は、シミュレーションや統計的な母数推定に利用される乱数の生成と検定の問題について考える。

乱数は一様分布を元にして、様々な確率分布に対して個別に求めることもできるが [1)]、任意の密度関数に対しても、マルコフ連鎖モンテカルロ法（以後 MCMC と呼ぶ）を用いることによって、求めることが可能である [2)]。我々は前者の乱数生成法について、元々College Analysis に含まれていたものを整理し、新たに MCMC による方法を追加して様々な応用分野で利用できるようにプログラムを強化した。その際、乱数の信頼性についても検討できるように、１次元乱数についてはその検定法を追加した。この検定はよく知られた分布だけでなく、一般的な密度関数（離散的な場合も含む）に対応できるようにした。

我々は例題などに利用するデータを作成するために、データ生成のプログラムを作成していた。その中には、同一データ、単調（定数）増加・減少、多項乱数の他に、１変量乱数が数種類と、よく利用される２変量正規乱数が含まれていた。我々はその中の１変量乱数について、これまでのフォーム上に貼り付けられたラジオボタン形式の固定メニューをコンボボックスに変更し、今後多くの乱数を登録できるようにした。

生成法が分かっている乱数以外の任意の乱数については、MCMC による乱数生成を利用する。これは生成したい乱数の密度関数をテキストボックスに入力して、生成することが可能であるが、あくまで近似的な方法なので、生成された乱数について十分な吟味が必要である。特にデータ生成プログラムに含まれる乱数については、そちらのプログラムを利用する方が精度は高い。詳細は３章で説明する。

乱数の検定は、密度関数をメニューの中から選択するか、テキストボックに直接密度関数を書き込んで実行する。検定の種類は、度数分布表やヒストグラムを理論分布と比較して見る方法、p-p プロット、p-p プロットを数値的に見るコルモゴロフースミルノフ検定、度数分布表に基づく適合度検定である。乱数の生成時にはこれらの検定を利用して精度を検討しておくことを勧める。

２．データの生成

データ生成についてはこれまで「ツール」の中で取り上げ、多くのサンプルの作成に使用してきたが、今回 MCMC や乱数の検定プログラムを作成するに当たり、分析の追加が容易なように変更した。メニュー［ツールーデータ生成］を選択すると、図 2.1 のような実行メニューが表示される。

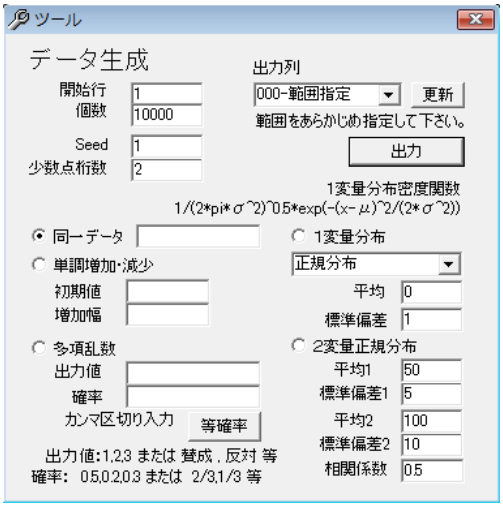


図 2.1　実行メニュー

データ生成は、予めグリッドを用意して実行する。最初に出力する列を指定するが、グリッドエディタのある列を選択して、そこへ出力する場合は、右上の「出力列」コンボボックスを「範囲指定」にする。その他には、「新規追加」で新しい列を追加して出力することもできるし、変数名から直接出力する列を選択することもできる。出力するデータをある行の範囲に限りたい場合、「開始行」と「個数」テキストボックスで出力する位置を指定することができる。乱数生成は、Seed で出力する乱数系列を指定するが、ここでも「Seed」テキストボックスに適当な数値を入れて、乱数の再現が可能になっている。出力する乱数の桁数は、「小数点桁数」テキストボックスに小数点以下の桁数を入れて指定する。どのような乱数を出力するかは、乱数の種類をラジオボタンで選択するが、「1 変量分布」については種類が多く、今後も追加する可能性があるので、コンボボックスから選択するようになっている。その際、選ぶ乱数によって、コンボボックスの上にある密度関数の表式と下にあるパラメータ入力用のテキストボックスの表示が変わる。図 2.2 にその例を示す。

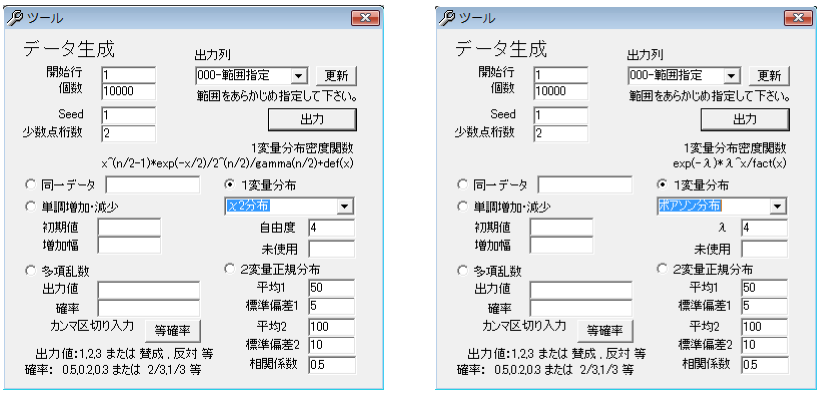


図 2.2　1 変量分布の変更

乱数出力が可能な 1 変量分布には、正規分布、対数正規分布、χ^2分布、F 分布、 t 分布、ガンマ分布、逆ガンマ分布、ベータ分布、ワイブル分布、指数分布、ポアソン分布、２項分布がある。これら（及びその他多く）の乱数の出力法については、参考文献 1）で詳細に与えられている。これら以外でメニューに表示されている出力値として、同一データ、単調増加（減少）、多項乱数、2 変量正規分布があるが、「同一データ」は、数値だけでなく文字列も出力可能である。また、「単調増加（減少）」は、初期値と増加（減少）幅を与えて 1 次関数的に変化するデータを出力する。「多項乱数」は、カンマ区切りで指定した出力値をその下のカンマ区切りで与えた確率で出力する。「2 変量正規分布」は、２つの変数の平均と標準偏差、相関係数を与えて乱数を出力する。乱数生成は設定を終えた上で「出力」ボタンをクリックする。

３．マルコフ連鎖モンテカルロ法

マルコフ連鎖モンテカルロ法（MCMC）は共分散構造分析やベイズ統計などで母数推定の有力な手法として利用される。我々は、その性質を調べるために、MCMC を利用した乱数生成のプログラムを作成した。生成された乱数はヒストグラムで表示され、理論分布と比較することができ、そのままデータとしてグリッドに出力することができる。ここでは最初にマルコフ連鎖モンテカルロ法の理論について述べ、次にプログラムの利用法について説明する。

3.1 マルコフ連鎖モンテカルロ法の理論

最初に我々は参考文献 2）に沿って、マルコフ連鎖モンテカルロ法についてまとめておくことにする。時刻 t に値 x が確率 $\pi^{(t)}(x)$ で生じる、ある確率変数 X について、この値が、時刻 t と共に変化して行く過程 $x^{(1)}, x^{(2)}, \cdots, x^{(t)}, \cdots$ を確率過程という。マルコフ連鎖は、この確率過程が時刻 t まで実現した後に、時刻 $t+1$ での値 $x^{(t+1)}$ の生成確率 $P(X = x^{(t+1)} | x^{(1)}, \cdots, x^{(t)})$ が時刻 t の値 $x^{(t)}$ だけによって決まるものをいう。すなわち、

$$P(X = x^{(t+1)} | x^{(1)}, \cdots, x^{(t)}) = P(X = x^{(t+1)} \mid x^{(t)})$$

である。

$$p(x^{(t+1)} \mid x^{(t)}) \equiv P(X = x^{(t+1)} \mid x^{(t)})$$

とすると、この $p(x^{(t+1)} \mid x^{(t)})$ は推移核と呼ばれる。値が離散的で有限個の場合、推移核はある有限な定数行列（推移行列）となる。マルコフ連鎖が既約的、正回帰的、かつ非周期的であるとき、エルゴード的であると言われ、以下の性質を満たすことが知られている。

$$\lim_{t \to \infty} \pi^{(t)}(x) = \pi(x)$$

ここに $\pi(x)$ はある不変分布である。即ち、どの状態から出発しても、$t \to \infty$ ではある状態 $\pi(x)$ に収束する。この状態を利用すると、以下の関係が成り立つことが分かる。

$$\pi(x^{(t+1)}) = \int \pi(x^{(t)})p(x^{(t+1)}|x^{(t)})dx^{(t)}$$

マルコフ連鎖が不変分布になっているための十分条件は隣接する 2 つの時刻 $t, t+1$ に対して以下の詳細つり合い条件が成り立つことである。

$$\pi\left(x^{(t)}\right)p\left(x^{(t+1)}\middle|x^{(t)}\right) = \pi\left(x^{(t+1)}\right)p\left(x^{(t)}\middle|x^{(t+1)}\right)$$

我々はある提案分布により乱数を生成し、ある条件に従ってこの詳細つり合い条件を満たすようにデータをサンプリングする。我々の提案分布の密度関数を$q(x_1|x_2)$とすると、通常この分布は詳細つり合い条件を満たさない。

$$\pi\left(x^{(t)}\right)q\left(x^{(t+1)}\middle|x^{(t)}\right) \neq \pi\left(x^{(t+1)}\right)q\left(x^{(t)}\middle|x^{(t+1)}\right)$$

さて、ここで、推移核 $p(x|x')$ をこの提案分布確率密度 $q(x|x')$ と、ある確率 $\alpha(x|x')$ を用いて以下のように表す。

$$p(x|x') = cq(x|x')\,\alpha(x|x')$$

ここに c は定数である。これは提案分布によって生成した乱数を確率$\alpha(x|x')$で選別して推移核の定数倍に一致させようとするものである。

この関係を詳細つり合い条件に代入すると定数 c の自由度を残して以下となる。

$$\pi\left(x^{(t)}\right)q\left(x^{(t+1)}\middle|x^{(t)}\right)\,\alpha\left(x^{(t+1)}\middle|x^{(t)}\right) = \pi\left(x^{(t+1)}\right)q\left(x^{(t)}\middle|x^{(t+1)}\right)\,\alpha\left(x^{(t)}\middle|x^{(t+1)}\right)$$

確率の$\alpha(x|x')$値は 0 から 1 の範囲で、以下のように決めれば良いことが分かる。

$\pi\left(x^{(t)}\right)q\left(x^{(t+1)}\middle|x^{(t)}\right) = \pi\left(x^{(t+1)}\right)q\left(x^{(t)}\middle|x^{(t+1)}\right)$ のとき、

$$\alpha\left(x^{(t+1)}\middle|x^{(t)}\right) = 1\,, \qquad \alpha\left(x^{(t)}\middle|x^{(t+1)}\right) = 1$$

$0 \leq \pi\left(x^{(t)}\right)q\left(x^{(t+1)}\middle|x^{(t)}\right) < \pi\left(x^{(t+1)}\right)q\left(x^{(t)}\middle|x^{(t+1)}\right)$ のとき、

$$\alpha\left(x^{(t+1)}\middle|x^{(t)}\right) = 1\,, \qquad \alpha\left(x^{(t)}\middle|x^{(t+1)}\right) = \frac{\pi\left(x^{(t)}\right)q\left(x^{(t+1)}\middle|x^{(t)}\right)}{\pi(x^{(t+1)})q\left(x^{(t)}\middle|x^{(t+1)}\right)} < 1$$

$\pi\left(x^{(t)}\right)q\left(x^{(t+1)}\middle|x^{(t)}\right) > \pi\left(x^{(t+1)}\right)q\left(x^{(t)}\middle|x^{(t+1)}\right) \geq 0$ のとき、

$$\alpha\left(x^{(t+1)}\middle|x^{(t)}\right) = \frac{\pi\left(x^{(t+1)}\right)q\left(x^{(t)}\middle|x^{(t+1)}\right)}{\pi(x^{(t)})q\left(x^{(t+1)}\middle|x^{(t)}\right)} < 1\,, \qquad \alpha\left(x^{(t)}\middle|x^{(t+1)}\right) = 1$$

これを $\alpha\left(x^{(t+1)}\middle|x^{(t)}\right)$ についてまとめると以下となる。

$$\alpha\left(x^{(t+1)}\middle|x^{(t)}\right) = \begin{cases} \min\left[\dfrac{\pi\left(x^{(t+1)}\right)q\left(x^{(t)}\middle|x^{(t+1)}\right)}{\pi(x^{(t)})q\left(x^{(t+1)}\middle|x^{(t)}\right)}, 1\right] & \text{分母} > 0 \\ 1 & \text{分母} = 0 \end{cases}$$

即ち、乱数を提案分布により生成し、確率 $\alpha\left(x^{(t+1)}\middle|x^{(t)}\right)$ によって抽出すれば、目的の分布に従う乱数を得ることができる。このような乱数生成のアルゴリズムを Metropolis-Hastings アルゴリズムという。

さて、任意の密度関数 $\pi(x)$ からの乱数を得るために、提案分布として我々のプログラムでは正規分布を考える。その確率密度関数は以下である。

$$q(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

この乱数の生成法について、酔歩的に前時刻の位置を中心として生成する場合と前回とは全く独立に生成する場合を考える。前者を酔歩連鎖、後者を独立連鎖と呼ぶ。

酔歩連鎖では、状態 x' から状態 x への推移は、x' を中心として上の正規分布を生成するので、$q(x|x') = q(x - x')$ となり、条件付き確率は具体的に以下となる。

$$q\left(x^{(t)}|x^{(t+1)}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(x^{(t)}-x^{(t+1)}-\mu\right)^2}{2\sigma^2}}$$

$$q\left(x^{(t+1)}|x^{(t)}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(x^{(t+1)}-x^{(t)}-\mu\right)^2}{2\sigma^2}}$$

ここで、 $\mu = 0$ の場合は $q\left(x^{(t)}|x^{(t+1)}\right) = q\left(x^{(t+1)}|x^{(t)}\right)$ となることから、確率を決める式は以下となる。

$$\frac{\pi\left(x^{(t+1)}\right)q\left(x^{(t)}\middle|x^{(t+1)}\right)}{\pi(x^{(t)})q\left(x^{(t+1)}\middle|x^{(t)}\right)} = \frac{\pi\left(x^{(t+1)}\right)}{\pi(x^{(t)})}$$

次に独立連鎖の場合は、これまでの位置に関係なく、上の乱数を生成するので、

$$q\left(x^{(t)}|x^{(t+1)}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(x^{(t)}-\mu\right)^2}{2\sigma^2}}$$

$$q\left(x^{(t+1)}|x^{(t)}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(x^{(t+1)}-\mu\right)^2}{2\sigma^2}}$$

となり、確率を決める式は以下となる

$$\frac{\pi\left(x^{(t+1)}\right)q\left(x^{(t)}\middle|x^{(t+1)}\right)}{\pi(x^{(t)})q\left(x^{(t+1)}\middle|x^{(t)}\right)} = \frac{\pi\left(x^{(t+1)}\right)e^{-\frac{\left(x^{(t)}-\mu\right)^2}{2\sigma^2}}}{\pi(x^{(t)})e^{-\frac{\left(x^{(t+1)}-\mu\right)^2}{2\sigma^2}}}$$

この関係は、離散分布の場合にも適用され、我々は正規分布から得られた値を、小数点以下 1 桁目の四捨五入により整数化して、提案分布として利用している。

次にこれを変数が複数ある場合に拡張する。時系列データを $x_i^{(t)}$ とし、提案分布として我々は独立な正規分布を考える。

$$q(x_1, \cdots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}}$$

n 変数の場合も、1 変数の場合と同様に、酔歩連鎖と独立連鎖を考える。特に酔歩連鎖では $\mu_i = 0\ (i = 1, \cdots, n)$ とする。

提案分布からの抽出確率は以下となる。

$$\alpha\left(x_i^{(t+1)} \middle| x_1^{(t+1)}, \cdots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \cdots, x_n^{(t)}\right)$$

$$= \begin{cases} min\left[\dfrac{\pi\left(\cdots, x_i^{(t+1)}, x_{i+1}^{(t)}, \cdots\right) q\left(x_i^{(t)} \middle| \cdots, x_i^{(t+1)}, x_{i+1}^{(t)}, \cdots\right)}{\pi\left(\cdots, x_{i-1}^{(t+1)}, x_i^{(t)}, \cdots\right) q\left(x_i^{(t+1)} \middle| \cdots, x_{i-1}^{(t+1)}, x_i^{(t)}, \cdots\right)}, 1\right] & \text{分母} > 0 \\ 1 & \text{分母} = 0 \end{cases}$$

ここで、変数の順番を変えて次の時点の乱数を求めたとしても、抽出された乱数の分布には影響がないことが知られている。

具体的に提案分布として上の独立な正規分布を考えると、酔歩連鎖の場合、

$$q\left(x_i^{(t+1)} \middle| x_1^{(t+1)}, \cdots, x_{i-1}^{(t+1)}, x_i^{(t)}, \cdots, x_n^{(t)}\right)$$

$$= \prod_{j=1}^{i-1} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{x_j^{(t+1)\,2}}{2\sigma_j^2}} \times \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{\left(x_i^{(t+1)} - x_i^{(t)}\right)^2}{2\sigma_i^2}} \times \prod_{k=i+1}^{n} \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{x_k^{(t)\,2}}{2\sigma_k^2}}$$

$$= q\left(x_i^{(t)} \middle| x_1^{(t+1)}, \cdots, x_i^{(t+1)}, x_{i+1}^{(t)}, \cdots, x_n^{(t)}\right)$$

より、以下となる。

$$\alpha\left(x_i^{(t+1)} \middle| x_1^{(t+1)}, \cdots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \cdots, x_n^{(t)}\right)$$

$$= \begin{cases} min\left[\dfrac{\pi\left(x_1^{(t+1)}, \cdots, x_i^{(t+1)}, x_{i+1}^{(t)}, \cdots, x_n^{(t)}\right)}{\pi\left(x_1^{(t+1)}, \cdots, x_{i-1}^{(t+1)}, x_i^{(t)}, \cdots, x_n^{(t)}\right)}, 1\right] & \text{分母} > 0 \\ 1 & \text{分母} = 0 \end{cases}$$

独立連鎖の場合も同様に求めることができるので省略する。

3.2 プログラムの利用法

メニュー［分析－基本統計－MCMC 乱数生成］を選択すると、図 3.1 のようなメニューが表示される。

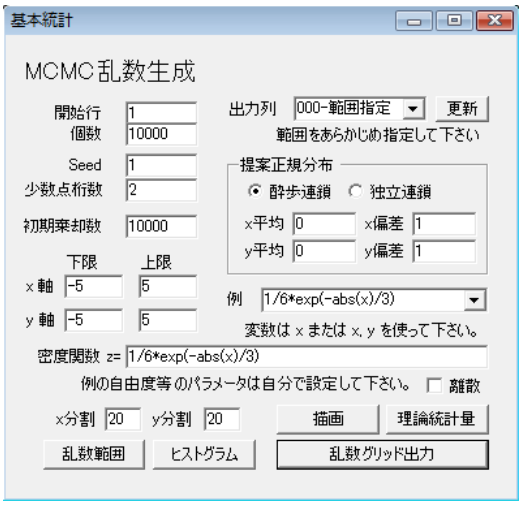


図 3.1　MCMC 乱数生成メニュー

プログラムを利用する際、まず「密度関数」テキストボックスに、出力したい乱数の密度関数を入力する。「例」のコンボボックスにサンプルが入っているので、それを参考にしてもらってもよい。ここではまず、密度関数 = 1/6*exp(-abs(x)/3) の 1 次元分布の例を用いて説明を行う。

目的分布の密度関数を入力した後、描画範囲の x 軸の上限と下限を入力する。この範囲はあくまで描画する際の表示範囲で、乱数生成はこれにとらわれない。乱数の生成範囲は、「乱数範囲」ボタンで、図 3.2 のように表示される。描画範囲が不明の場合はこの結果を参考にしてもよい。

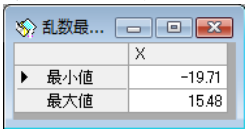


図 3.2　乱数生成の範囲

描画範囲として下限-20 と上限 20 を入力したら、まず、「ヒストグラム」ボタンで図 3.3a のようなヒストグラムを描いてみる。

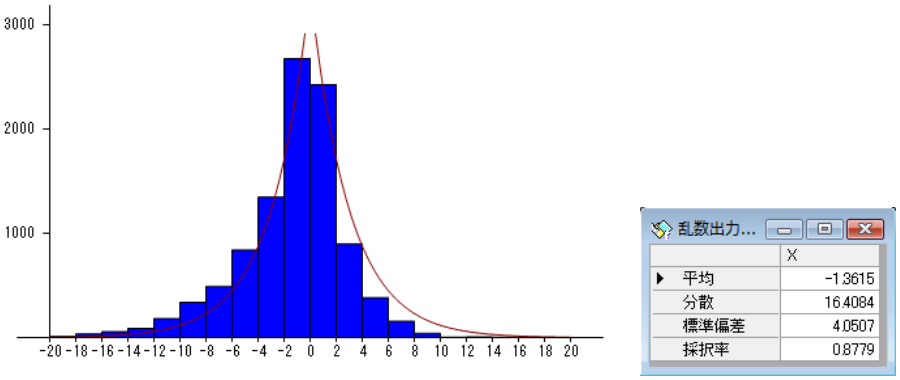


図 3.3a　乱数のヒストグラムと理論曲線（Seed=1）

ヒストグラムと同時に、出力した乱数の統計量も表示される。採択率は、Metropolis-Hastings アルゴリズムの抽出率をいう。

図 3.3a の中の曲線は目的分布の密度関数を利用した理論値である。この場合少しずれているが、乱数の「Seed」を変えることによって分布が異なってくる。例として、図 3.3b に Seed = 2 の場合を示す。

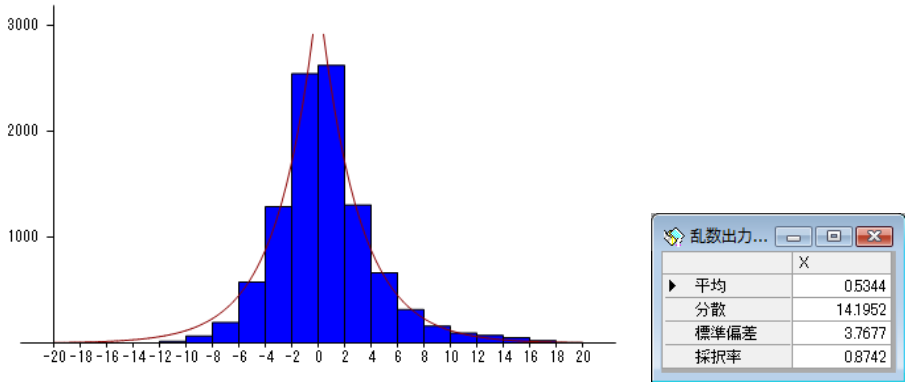


図 3.3b　乱数のヒストグラムと理論曲線（Seed=2）

ヒストグラムの階級幅は「x 分割」の数によって決まる。この場合、上限と下限の差が 40 で x 分割数が 20 であるので階級幅は 2 になっている。

密度関数の形は、「描画」ボタンで見ることができる。ここでは 1 変量関数グラフのプログラムを利用するので、そのメニューが表示されるが、その中の「グラフ描画」ボタンをクリックすると図 3.4 のようなグラフが表示される。

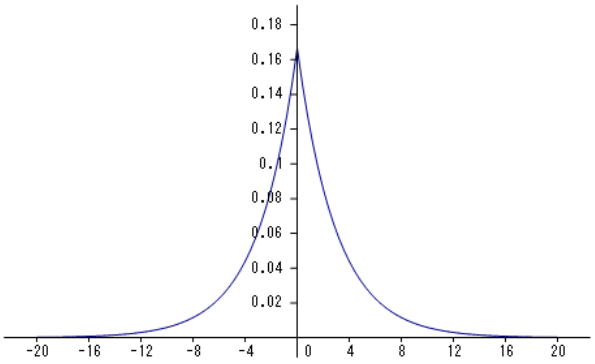


図 3.4　密度関数グラフ

密度関数から求められる、平均、分散、標準偏差は、「理論統計量」ボタンで図 3.5 のように表示される。

図 3.5　理論統計量結果

目的分布の関数形のみ分かって、規格化定数が不明の場合は、定数の部分に表示された値（1／面積）を掛けておけばよい。乱数生成は規格化定数にはよらないので、特に掛けておく必要もない。

提案分布については、酔歩乱数の場合、平均は 0 とし、標準偏差は目的分布のものより小さくしておくと無難である。提案分布の標準偏差を大きくして行くと乱数の尖度が小さくなる傾向があるので、適当な標準偏差を選ぶことは重要である。また独立連鎖の場合、提案分布の平均と標準偏差を目的分布に合わせておくと無難である。

以上のようにして求めた乱数は、データとしてグリッドに出力できる。予め求めたい乱数の数の行数を持ったグリッドを用意しておき、「出力列」コンボボックスで「範囲指定」を選び、列を選択して、「乱数グリッド出力」ボタンをクリックする。また、「出力列」で「新規追加」を選択すると、新しい列を追加して乱数を出力する。これは、メニュー「ツールーデータ生成」の乱数生成と同じである。

次に離散的な乱数生成について説明する。例えば「例」で、ポアソン分布を選択すると、「密度関数」テキストボックスには、密度関数＝ exp(-λ)*λ^x/fact(x) が表示され、右下の「離散」チェックボックスにチェックが入る。離散分布の場合は、この「離散」チェックボックスのチェックが重要である。密度関数にはパラメータλが含まれているが、利用者はこれを書き換えて適当な値にする。例えば、λを 3 とすると、exp(-3)*3^x/fact(x) となる。生成した乱数の最小値と最大値は「乱数範囲」ボタンをクリックすることにより、0 と 9 であるから、「下限」を 0、「上限」を 10 にして、「ヒストグラム」ボタンをクリックすると図 3.6 のようになる。

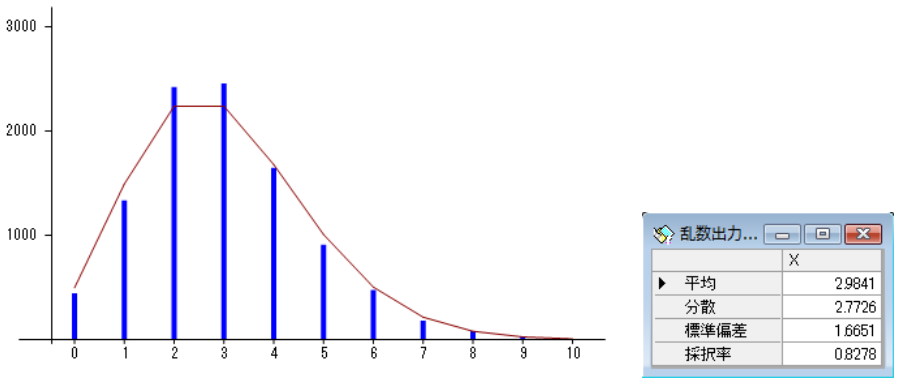


図 3.6　ポアソン分布

現在のバージョンでは、離散分布は 1 次元の場合だけに対応している。また、「描画」ボタンは離散分布に対応していない。

次に 2 次元の分布について見る。変数は x と y で与える。例として、密度関数のコンボボックスで 2 変量正規分布を選ぶと、以下のような 2 変量正規分布の密度関数の式が表示される。

z = 1/(2*pi*(1-r^2)^0.5)*exp(-(x^2-2*r*x*y+y^2)/2/(1-r^2))

ここで、r は相関係数を表す。例えば r を 0.5 と書き換えて、「描画」ボタンをクリックし、表示された 2 変量関数グラフのメニューで、そのまま「グラフ描画」ボタンをクリックすると、図 3.7 のような密度関数グラフが表示される。

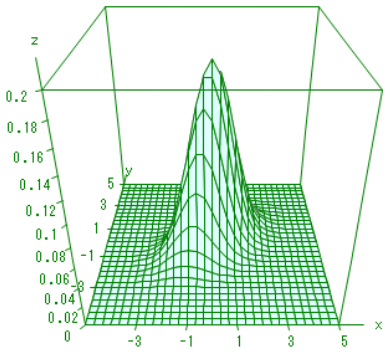


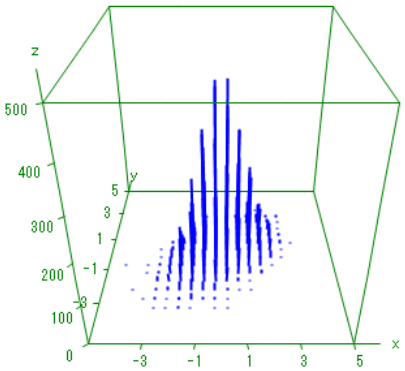
図 3.7　2 変量正規分布密度関数

次に、「統計量」ボタンをクリックすると、図 3.8 に示されるような結果が表示される。

面積(s)	1.0000
定数(1/s)	1.0000
X平均	0.0000
X分散	1.0000
X標準偏差	1.0000
Y平均	0.0000
Y分散	1.0000
Y標準偏差	1.0000
相関係数	0.5000

図 3.8　統計量結果

出力される乱数の分布を見るために「ヒストグラム」ボタンをクリックすると図 3.9 のような 2 変量ヒストグラムが表示される。棒グラフは見にくいことと描きにくいことから、度数を太めの直線で表わしている。



乱数出力結果

	X	Y
平均	0.0137	0.0115
分散	1.0023	0.9579
標準偏差	1.0012	0.9787
採択率	0.6672	0.6620

図 3.9　2 変量ヒストグラム

2 変量の場合のグリッドへの乱数出力は、2 列同時に出力されるので注意を要する。

４．分布の検定

乱数データが与えられている場合、それが本当に自分が求める分布に従っているかどうか調べることは重要である。ここではこの分布の検定法について説明する。College Analysis でメニュー［分析－基本統計－分布の検定］を選択すると図 4.1 のような分析メニューが表示される。

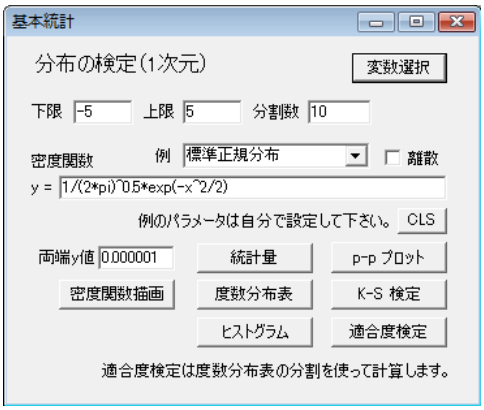


図 4.1　分析メニュー

データは縦 1 列でグリッドエディタに入力されたものを使う。「変数選択」で、検定するデータの変数を 1 つ選択し、メニューの「y =」テキストボックスに密度関数の形を数式で入力する。よく知られた分布の場合は、上の「例」コンボボックスから図 4.2a のように選び、図 4.2b のようにパラメータと「下限」、「上限」を変更する。ここでは、自由度 3 の χ^2 分布を例にする。

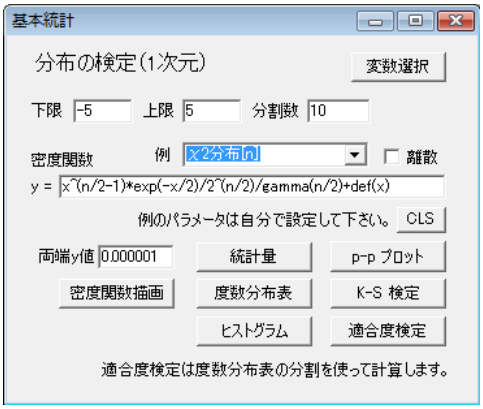


図 4.2a　密度関数の指定

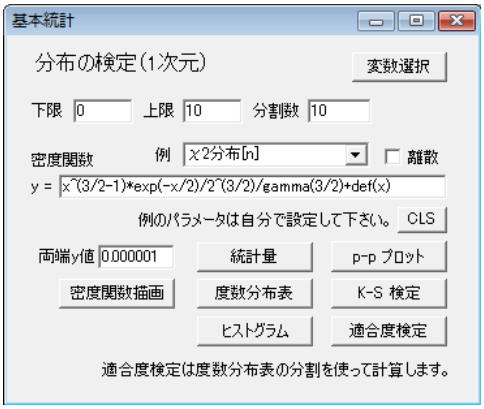


図 4.2b　パラメータと下限・上限の指定

密度関数の性質を見るために、「統計量」ボタンをクリックすると図 4.3 の結果を得る。

	データ	理論値
最小(全確率)	0.0300	1.0000
最大(1/全確率)	15.0900	1.0000
平均	3.0830	3.0000
分散	6.2460	6.0000
標準偏差	2.4992	2.4495

図 4.3　統計量

これはデータを用いた統計量と統計量の理論値との比較である。但し、最小（全確率）と最大（1/全確率）は、データでは最小と最大、理論値では全確率と 1/全確率を表す。

次に「度数分布表」ボタンをクリックするとデータと理論値の度数分布の比較が、図 4.4 のように表示される。

	度数	比率	理論度数	理論比率
領域なし	0	0.0000	0.00	0.0000
0.0<=x<1.0	194	0.1940	198.72	0.1987
1.0<=x<2.0	216	0.2160	228.85	0.2288
2.0<=x<3.0	177	0.1770	180.78	0.1808
3.0<=x<4.0	134	0.1340	130.16	0.1302
4.0<=x<5.0	106	0.1060	89.67	0.0897
5.0<=x<6.0	63	0.0630	60.19	0.0602
6.0<=x<7.0	32	0.0320	39.71	0.0397
7.0<=x<8.0	28	0.0280	25.89	0.0259
8.0<=x<9.0	17	0.0170	16.72	0.0167
9.0<=x<10.0	11	0.0110	10.72	0.0107
10.0<=x<30.0	22	0.0220	18.56	0.0186
合計	1000	1.0000	999.97	1.0000

図 4.4　連続分布の度数分布表

合計を除く一番上と一番下は、「下限」と「上限」に指定された領域以外についての度数と比率の和である。ここで領域外の範囲は、密度関数の高さが分析メニューの「両端 y 値」で指定された値より小さくなった点までを計算する。図 4.4 では「10.0<=x<30.0」の 30.0 がその点である。

次に、分析メニューで「ヒストグラム」をクリックすると、上の度数分布表の「下限」と「上限」の範囲内のデータと理論的な密度曲線が図 4.5 のように表示される。

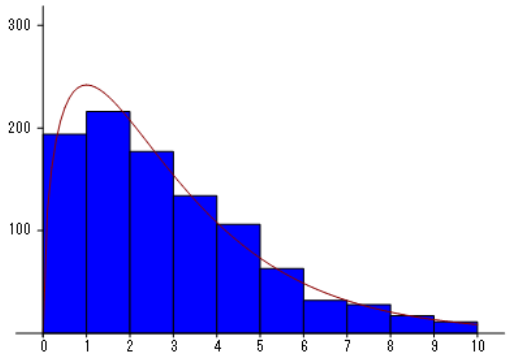


図 4.5　連続分布のヒストグラム

度数分布表やヒストグラムにより、定性的な分布の検討ができる。

次にもう少し、分布との一致を見易くするために、分析メニューの「p-p プロット」をクリックする。結果は図 4.6 のようになる。

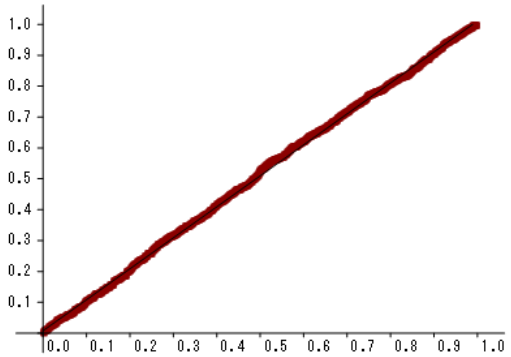


図 4.6　p-p プロット

これは、データと理論値の適合性を見るための直線で、適合が良ければプロットはこの図のように直線状に並ぶ。これは正規性の検定の「正規確率紙」の方法（一般に q-q プロットと呼ぶ）に類似するものである。

p-p プロットを数値的に検定する方法がコルモゴロフースミルノフ（Kolmogorov-Smirnov）検定である。これは略して、K-S 検定と呼ばれる。この検定はプロットがこの直線から最大どれ位離れているかで適合の検定確率を求める。分析メニューで「K-S 検定」ボタンをクリックすると図 4.7 のような結果が得られる。

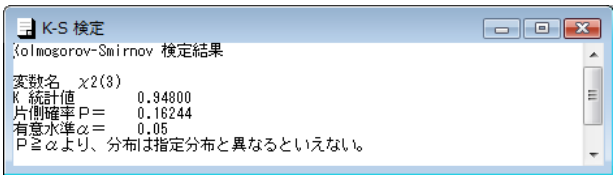


図 4.7　K-S 検定結果

また分布の検定には、図 4.4 の度数分布表をもとに、度数分布が理論比率に合っているかどうかを調べる適合度検定がある。これは分析メニューの「適合度検定」ボタンをクリックして得られる。分割は、度数分布表で与えられる分割を利用する。但し、理論比率が 0 の部分は分析から除外する。結果を図 4.8 に示す。

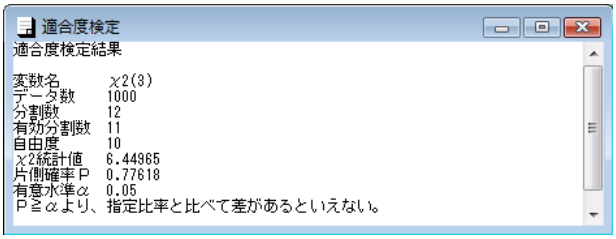


図 4.8　適合度検定結果

この適合度検定は離散的な分布に対しても適用できる。分析メニューの離散チェックボックスにチェックを入れた後に「度数分布表」ボタンをクリックして表示される、λ=4 のポアソン分布に対する度数分布表を図 4.9 に示す。

	度数	比率	理論度数	理論比率
-1<=x<=-1	0	0.0000	0.00	0.0000
x=0	22	0.0220	18.32	0.0183
x=1	60	0.0600	73.26	0.0733
x=2	142	0.1420	146.53	0.1465
x=3	179	0.1790	195.37	0.1954
x=4	221	0.2210	195.37	0.1954
x=5	156	0.1560	156.29	0.1563
x=6	97	0.0970	104.20	0.1042
x=7	55	0.0550	59.54	0.0595
x=8	37	0.0370	29.77	0.0298
x=9	19	0.0190	13.23	0.0132
x=10	8	0.0080	5.29	0.0053
11<=x<=17	4	0.0040	2.84	0.0028
合計	1000	1.0000	1000.00	1.0000

図 4.9　離散分布の度数分布表

これを「ヒストグラム」で表わすと図 4.10 のようになる。

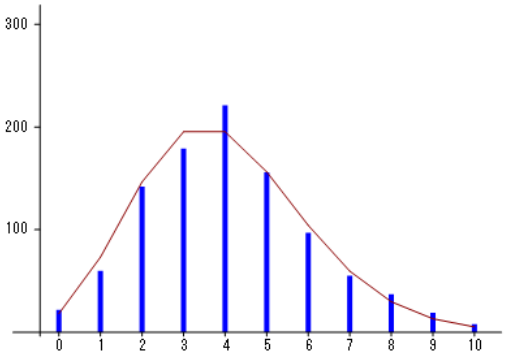


図 4.10　離散分布のヒストグラム

この乱数について「適合度検定」を実行すると図 4.11 のような結果となる。

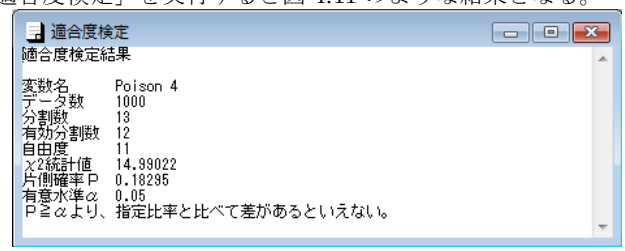


図 4.11　適合度検定結果

最後に、連続分布の場合は、「密度関数描画」ボタンで、関数描画用のメニューが表示され、関数グラフを描くことができる。

仮説検定を利用する場合、検定結果から、分布と異なることは示されるが、指定された分布になるという保証はない。特に、データ数が少ない場合には、有意差を見出すことが困難なため、注意を要する。また、連続分布の場合、分割数をいくつにするのか、どこに分割の境界を持ってくるのかで、検定結果が変わる場合もある。いろいろな場合で試して、総合的に確信を得る以外に方法はないのではなかろうか。

５．おわりに

我々は、シミュレーションや統計的な母数推定に利用される乱数生成の問題についてプログラムを作成した。これまでは、メニュー［ツールーデータ生成］（元は「データ発生」となっていた）の中で、よく知られた確率分布に対して、分布に応じた方法で乱数生成を行っていたが、分布の数も少なく、分析メニューも固定的であった。今回はこれを変更し、特に重要な１変量分布については、コンボボックスから選ぶことにして、容易に分布の種類を増やせるようにした。またこれを使って、新たな分布も追加した。現在のバージョンでは、正規分布、χ^2 分布、Ｆ分布、ｔ分布、対数正規分布、ガンマ分布、逆ガンマ分布、ベータ分布、ワイブル分布、指数分布、ポアソン分布、二項分布が含ま

れているが、今後必要が生じた場合は追加して行く予定である。これらの乱数については精度良く生成される。

これら以外の乱数や自分で密度関数を定義する乱数については、MCMC を利用して求める。MCMC は最初に生成したデータを大量に破棄したり、生成したデータを棄却したりしながら、乱数を求めるので、生成効率は良いとは言えず、分布の適合度にも問題が残るが、任意の密度関数に対応する乱数を容易に求めることができるという魅力もある。その本質的な部分は、我々のプログラムの中でわずか 100 行余りである。このように乱数生成法は種類や用途に応じて使い分ける必要がある。

乱数の生成で問題となるのは、生成した乱数の正当性である。このプログラムでは、基本的な方法である、ヒストグラム、p-p プロット、適合度検定、コルモゴロフースミルノフ検定を取り上げた。これらの方法を利用して、上の 2 つの乱数生成法を比較すると、前者の方法がはるかに高い精度をもっている。MCMC はヒストグラムで見ると良さそうに思われるが、数値的にはなかなか良い結果が得られない。特に自由度が小さい χ^2 分布などの 0 の近くの立ち上がりには問題があり、提案分布の選び方にも大きく左右される。現在は棄却数 10,000、データ数 1,000～10,000、酔歩連鎖で平均 0、標準偏差を指定分布の標準偏差より少し小さくして試しているが、今後設定を変えて検討し、効率の良い方法を見付けて行かなければならない。MCMC の場合、理論を理解することと実際に正しい乱数を生成することとの間にはかなり差があるようである。

参考文献

1) 計算機シミュレーションのための確率分布乱数生成法, 四辻哲章, プレアデス出版, 2010.

2) マルコフ連鎖モンテカルロ法, 豊田秀樹, 朝倉書店, 2008.

Multi-purpose Program for Social System Analysis 21

- Random Number Generation and Testing -

Masayasu FUKUI, Hong Yan MENG, Meng WU and Yong Jie CUI

Department of Business Administration, Faculty of Business Administration,
Fukuyama Heisei University

Abstract

We have been constructing a unified program on the social system analysis for the purpose of education. This time, we created programs of random number generation which is used in simulation or in statistical parameter estimation. We also made a program on the test of fitting to the distribution. In the random number generation, we used specific method to the distribution and Markov chain Monte Carlo method. In the test of fitting, histogram, p-p plot, goodness of fit test and Kolmogorov - Smirnov test are included in our program.

Keywords

College Analysis, social system analysis, statistics, random number, Markov chain Monte Carlo methods, MCMC

URL: http://www.heisei-u.ac.jp/ba/fukui/