炉物理プログラム演習：乱数の発生と確率分布の計算

ある変数を考えたとき、それがとりうる値に対して確率が与えられている場合には、その変数を**確率変数**という。可算集合 $\{x_1, x_2, ...\}$ の中の値をとる確率変数は**離散型**と言われ、例えばサイコロの出目が挙げられるであろう。一方、確率変数 X のとる値が関数 $f(x)$ によって

$$P(a \leq X \leq b) = \int_a^b f(x)dx \tag{1}$$

のように表される場合、X は**連続型**と言われる。

計算機上でこのような確率変数を扱う場合、一般的には**擬似乱数**と呼ばれるものを利用する [1]。計算機での乱数の発生法には様々なものがあるが、ここでは最も基本的と言えるであろう C 言語の RAND 関数を利用する。RAND 関数は [0,RAND_MAX] の範囲の整数値を返すので、[0,1] の範囲の実数の乱数を必要とするときには、RAND 関数の戻り値を RAND_MAX で除してやればよいであろう。

RAND 関数は標準ライブラリ `stdlib.h` をインクルードすることにより利用でき、`rand()` もしくは `random()` メソッドにより [0,RAND_MAX] の範囲の整数値が返される。以下に RAND 関数を用いた乱数の発生例を示す。

Listing 1: RAND 関数を用いた乱数の発生例

```
1  #include <stdlib.h>
2  #include <iostream>
3
4  using namespace std;
5
6  int main()
7  {
8    srand(10);
9
10   cout<<RAND_MAX<<"\n";
11   cout<<rand()<<"\n";
12   cout<<random()<<"\n";
13
14   return 0;
15 };
```

2 行目は出力のための `cout` メソッドを利用するためのライブラリのインクルードを示し、4 行目は `std::`の記法を省略するためのものである。また、8 行目の `srand` メソッドは発生させる乱数列を変えるためのものであり、この引数を変えなければ実行のたびに同一の結果が得られ、これを変更することにより異なる結果を得ることができる（試してみるとよい）。

現在、非常に使い勝手が良く、周期の長い擬似乱数発生機能が C++に複数実装されているようであるが、今回は演習が目的なので、この RAND 関数を利用する課題を提示する。

はじめに、乱数の発生方法の演習として、以下の課題に取り組んでもらいたい。

問題１：乱数を用いて円周率を計算する。これは一辺の長さが 1 の正方形と直径が 1 の円の面積の比が 1:$\pi/4$ であることを利用する。すなわち、直径が 1 の円が内接する一辺の長さが 1 の正方形内にランダムに位置を決め、それが円内に存在する確率を計算すればよい。乱数の発生回数と評価された円周率の関係を求め、乱数の発生回数の増加に伴い、円周率の推定値が厳密値に近づいていくことを確認せよ。

次に、乱数を用いてある確率分布に従う確率変数の標本群を求め、それがどの程度、想定している確率変数を再現するか評価してみよう。

問題２：サイコロの出目の頻度分布を作成せよ。試行回数を 10、100、1,000、10,000 と増やしていったときに、得られた頻度分布が厳密な確率分布に一致していく様子を観察せよ。

今度は、複数のサイコロを同時に振ったときの、その出目の平均値の確率分布を考えよう。サイコロが１個のときは一様分布になることが容易に想像できるが、同時に振るサイコロの数が増えるに従って、この確率分布はどのように変化していくであろうか。

問題３：複数のサイコロの出目の平均値の頻度分布を作成せよ。試行回数は 10,000 回とする。なお、同時に振るサイコロの個数を 2 から 5 まで増加させ、頻度分布がどのように変化していくか観察せよ。

[1] ここで「擬似」としているのは、これが厳密に「乱れた（ランダムな）」変数ではなく、超長期的にはある種の周期性を持つことが理由である。

工学の問題でしばしば現れるのは、確率変数が正規分布に従うというものである。正規分布に従う確率変数についてのランダムな標本を得たい場合、擬似乱数は [0,1] の一様分布で発生するので、これを正規分布のものに変換する必要がある。このあたりの考え方については文献 [2] に詳しく記述されており、ここではその概要を述べる。

確率変数 X が $[a,b]$ の範囲で非ゼロの値をとる確率密度関数 $f(x)$ に従うとしたとき、X が $[a,x]$ の範囲の値をとる確率 $F(x)$ は以下のように定義できる。

$$F(x) = \int_a^x f(x')dx' \tag{2}$$

この $F(x)$ を**累積分布関数**と呼ぶ。$f(x)$ は確率密度を示す関数であるが、$F(x)$ は確率そのものを示す関数となる。定義から明らかなように、ここで示した例では $F(a) = 0$、$F(b) = 1$ であり $F(x)$ は単調増加関数となる。

このように確率密度関数 $f(x)$ が与えられたときにこれに従う確率変数（の標本）を乱数によりランダムに決定（抽出）する場合は、以下の方法により行う。まずは [0,1] の一様分布に従う確率変数からランダムに値を抽出する（これを ξ とする）。そして、以下の式から、$f(x)$ に従う確率変数 X の標本 $\tilde{x}$ を決める。

$$\xi = F(\tilde{x}) \tag{3}$$

この式は、F の逆関数 F^{-1} を用いて

$$\tilde{x} = F^{-1}(\xi) \tag{4}$$

と書ける。

一様分布に従う乱数を用いた正規分布に従う乱数の発生方法としては **Box-Muller 法**が広く知られている。これは以下の定理に基づく。

確率変数 X 及び Y が互いに独立で、ともに (0,1) 上での一様分布に従うものとする。このとき

$$Z_1 = \sqrt{-2\log X}\cos 2\pi Y, \tag{5}$$

$$Z_2 = \sqrt{-2\log X}\sin 2\pi Y \tag{6}$$

で定義される Z_1、Z_2 は、平均 0、分散 1 の標準正規分布に従う互いに独立な確率変数となる。

任意の平均 μ、分散 σ^2 の正規分布に従う確率変数の標本 $\tilde{z}$ を得たいときには、上述の方法で得た標準正規分布に従う確率変数の標本 z を

$$\tilde{z} = z\sigma + \mu \tag{7}$$

のように変換すればよい。

問題４：平均 5.38、標準偏差 0.55 の正規分布に従う確率変数を考える。乱数を用いてこの確率変数の標本を複数個作成し、それらに基づいて平均値、分散、歪度、尖度を計算せよ。そして、標本数をパラメータとして、これら統計量推定値を図示し、標本数が増加するにつれてこれら統計量がもとの確率分布（母分布）のものに一致することを示せ。

なお、上記の統計量推定値は、I 個の標本 x_i を用いて以下のように計算される。

$$\bar{x} = \sum_{i=1}^{I} x_i, \tag{8}$$

$$V = \sigma^2 = \frac{1}{I-1}\sum_{i=1}^{I}(x_i - \bar{x})^2, \tag{9}$$

$$\mathrm{Sk} = \frac{I}{(I-1)(I-2)}\sum_{I=1}^{I}\left(\frac{x_i - \bar{x}}{\sigma}\right)^3, \tag{10}$$

$$\mathrm{Ku} = \frac{I(I+1)}{(I-1)(I-2)(I-3)}\sum_{i=1}^{I}\frac{(x_i - \bar{x})^4}{\sigma^4} - \frac{3(I-1)^2}{(I-2)(I-3)} \tag{11}$$

ここで、$\bar{x}$、V、σ、Sk、Ku はそれぞれ平均、分散、標準偏差、歪度そして尖度の推定値に対応する。なお、ここで「推定値」としているのは、これらは平均 5.38、標準偏差 0.55 の正規分布（従って、歪度と尖度はともにゼロである）の母分布から取り出した有限個の標本に基づいて推定した統計量であるためである。従って、これらの統計量推定値は母分布の統計量とは一致しない（採取した標本数が無限大であれば一致する）。

なお、歪度、尖度については、標本分散の計算式に基づいた以下のような誤った式を用いてしまう場合があるので注意が必要である。

$$\mathrm{Sk} = \frac{1}{(I-1)}\sum_{I=1}^{I}\left(\frac{x_i-\bar{x}}{\sigma}\right)^3, \tag{12}$$

$$\mathrm{Ku} = \frac{1}{(I-1)}\sum_{i=1}^{I}\frac{(x_i-\bar{x})^4}{\sigma^4} - 3 \tag{13}$$

問題４において、この誤った式を用いた場合の結果も示すとよいであろう。

乱数を用いて抽出された確率変数の標本から推定した平均値や分散といった統計量は、それ自身が確率変数であり、それらに対しても平均、分散等の統計量が定義される。そこで、次は、標本から推定した平均値（標本平均）と分散（標本不偏分散）の不確かさ（分散）について考えよう [2] 。

ここで、確率変数 X、Y について、その線形結合 $aX+bY$ の分散 $V(aX+bY)$ は以下のように計算されることを記載しておく。

$$V(aX+bY) = a^2V(X) + b^2V(Y) \tag{14}$$

なお、$V(X)$、$V(Y)$ はそれぞれ X、Y の分散を示す。

確率変数 X の I 個の標本 x_i から得られる標本平均 $\bar{x} = \sum_{i=1}^{I} x_i/I$ の分散 $V(\bar{x})$ は以下のように求められる。

$$V(\bar{x}) = V\left(\sum_{i=1}^{I} x_i/I\right) = \frac{1}{I^2}\sum_{i=1}^{I}V(x_i) = \frac{1}{I^2}\sum_{i=1}^{I}V(X) = \frac{1}{I}V(X) \tag{15}$$

また、I 個の標本 x_i から得られる標本不偏分散 s^2 については、確率変数 X が正規分布に従うときには、

$$Z = (I-1)\,\frac{s^2}{V(X)} = \sum_{i=1}^{I}\frac{(x_i-\bar{x})^2}{V(X)} \tag{16}$$

で定義される確率変数 Z が自由度 $I-1$ のカイ二乗分布に従うことを利用すれば求めることが出来る。自由度 $I-1$ のカイ二乗分布の平均は $I-1$、分散は $2(I-1)$ になることから、$s^2 = \dfrac{V(X)}{I-1}Z$ より、標本不偏分散 s^2 の平均値として $V(X)$ が、分散として $\dfrac{2\{V(X)\}^2}{I-1}$ が、それぞれ得られる [3]。

問題５：問題４で求めた平均値（標本平均）、分散（標本不偏分散）について、それらの不確かさ（標準偏差）を上述の方法に基づいて評価し、それら標本から推定した統計量（問題４で求めた平均値、分散）ともとの確率分布（母分布）の統計量（X の平均値 $E(X)$ と分散 $V(X)$）とのずれが、推定した統計量の不確かさと整合していることを確認せよ。この評価はいくつかの異なる標本数で行えばよい。なお、標本平均、標本不偏分散の分散を求める際にはもとの確率分布の分散 $V(X)$ が必要となる。そこで、もとの確率分布の分散が分かっている（分散を 0.55^2）とした場合と、標本不偏分散で代用した場合とで結果を比較せよ。

問題６：着目する確率変数が正規分布に従うことが分かっていない場合は、標本不偏分散の不確かさを上述のように求めることは出来ない。このような場合は、有限個の標本に基づく標本不偏分散の計算を何度も繰り返し行うことにより、標本不偏分散の頻度分布を得て、そこから標本不偏分散の平均値（期待値）と分散を推定すればよい。問題４で求めた平均値と分散について、100 回の同様の計算を行い、その結果から各々の不確かさを評価せよ。

なお、標本不偏分散から計算される標準偏差 $s=\sqrt{s^2}$ の分散 $V(s)$ は以下の式を用いて計算できる [4]。

$$V(s^2) = \left(\frac{ds^2}{ds}\right)^2 V(s) = (2s)^2\,V(s) = 4s^2V(s) \tag{17}$$

[2] ここでの記述は文献 [3, 4] に拠る。

[3] $V(s^2) = V\left(\dfrac{V(X)}{I-1}Z\right) = \dfrac{(V(X))^2}{(I-1)^2}V(Z) = \dfrac{2\,(V(X))^2}{I-1}$

複数の確率変数についてランダムに標本を抽出する場合、それぞれが独立、すなわち相関係数がゼロであるならば、それぞれの確率変数について上記の方法で標本を計算すればよい。一方、確率変数間に相関がある場合には、それに対する配慮が必要である [4] 。

L 個の確率変数 $F_l(l=1,...,L)$ を考え、それぞれについてその確率分布に基づいて I 個の標本を得たとし、F_l の i 番目の標本を f_l^i と書く。F_l と $F_{l'}$ の標本共分散 $m_{ll'}$ は以下のように定義できる。

$$m_{ll'} = \frac{\sum_{i=1}^{I} \left(f_l^i - \bar{f}_l\right)\left(f_{l'}^i - \bar{f}_{l'}\right)}{I-1} \tag{18}$$

ここで $\bar{f}_l$ は F_l の標本平均を示す。

この複数の確率変数をまとめてベクトルで $\mathbf{f}$ と表記し、i 番目の標本及び標本平均をそれぞれ $\mathbf{f}^i$、$\bar{\mathbf{f}}$ と表記すると、$\mathbf{f}$ の共分散行列 $\mathbf{M}$ は以下のように記述できる。

$$\mathbf{M} = \left(\mathbf{F} - \bar{\mathbf{F}}\right)\left(\mathbf{F} - \bar{\mathbf{F}}\right)^T \tag{19}$$

ここで、$\mathbf{F}$、$\bar{\mathbf{F}}$ はいずれも $L \times I$ の行列で、

$$\mathbf{F} = \begin{pmatrix} \mathbf{f}^1 & \mathbf{f}^2 & \cdots & \mathbf{f}^I \end{pmatrix} = \begin{pmatrix} f_1^1 & f_1^2 & \cdots & f_1^I \\ f_2^1 & f_2^2 & \cdots & f_2^I \\ \vdots & \vdots & & \vdots \\ f_L^1 & f_L^2 & \cdots & f_L^I \end{pmatrix}, \tag{20}$$

$$\bar{\mathbf{F}} = \begin{pmatrix} \bar{\mathbf{f}} & \bar{\mathbf{f}} & \cdots & \bar{\mathbf{f}} \end{pmatrix} = \begin{pmatrix} \bar{f}_1 & \bar{f}_1 & \cdots & \bar{f}_1 \\ \bar{f}_2 & \bar{f}_2 & \cdots & \bar{f}_2 \\ \vdots & \vdots & & \vdots \\ \bar{f}_L & \bar{f}_L & \cdots & \bar{f}_L \end{pmatrix}, \tag{21}$$

である。

さて、次に $\mathbf{f}$ の線形変換 $\mathbf{f}'$ を以下のように定義する。

$$\mathbf{f}' = \mathbf{P}\mathbf{f} \tag{22}$$

式 (19) の両辺について、左から $\mathbf{P}$、右から $\mathbf{P}^T$ を乗じると以下の式を得る。

$$\begin{aligned} \mathbf{P}\mathbf{M}\mathbf{P}^T &= \left(\mathbf{P}\mathbf{F} - \mathbf{P}\bar{\mathbf{F}}\right)\left(\mathbf{F} - \bar{\mathbf{F}}\right)^T\mathbf{P}^T \\ &= \left(\mathbf{P}\mathbf{F} - \mathbf{P}\bar{\mathbf{F}}\right)\left(\mathbf{P}\left(\mathbf{F} - \bar{\mathbf{F}}\right)\right)^T \\ &= \left(\mathbf{P}\mathbf{F} - \mathbf{P}\bar{\mathbf{F}}\right)\left(\mathbf{P}\mathbf{F} - \mathbf{P}\bar{\mathbf{F}}\right)^T \\ &= \left(\mathbf{F}' - \bar{\mathbf{F}}'\right)\left(\mathbf{F}' - \bar{\mathbf{F}}'\right)^T = \mathbf{M}' \end{aligned} \tag{23}$$

つまり、行列 $\mathbf{P}\mathbf{M}\mathbf{P}^T$ は $\mathbf{f}'$ の共分散行列 $\mathbf{M}'$ となることが分かる。

ここで、線形変換行列 $\mathbf{P}$ を適切に選ぶことにより、$\mathbf{M}'$ を対角行列にすることが出来れば、すなわちパラメータ $\mathbf{f}'$ がそれぞれ独立関係にあるように出来れば、各々のパラメータ f_l' について前述の方法で標本を求め、以下の式で $\mathbf{f}$ に変換すればよいことが分かる。

$$\mathbf{f} = \mathbf{P}^{-1}\mathbf{f}' \tag{24}$$

次に、行列 $\mathbf{M}'$ を対角行列にするための方法として、ここでは特異値分解を用いた方法を説明する。

行列 $\mathbf{M}$ に特異値分解を施すと以下の式が得られる。

$$\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{U}\mathbf{D}\mathbf{U}^T \tag{25}$$

ここで、$\mathbf{M}$ は対称行列であることから、$\mathbf{U} = \mathbf{V}$ を用いている。また、$\mathbf{D}$ は対角行列である。$\mathbf{U}$ はユニタリ行列であるため、$\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}$ より、以下の式が得られる。

$$\mathbf{U}^T\mathbf{M}\mathbf{U} = \mathbf{D} \tag{26}$$

[4] 例えば、2 つの確率変数 A、B の間に強い正の相関がある場合、A の標本としてその平均値よりも大きい値が得られた場合、B の標本としてもその平均値より大きい値が得られるべきであると考えられる。

この式と式 (23) を比較することにより、線形変換行列として $\mathbf{P} = \mathbf{U}^T$ を用いることにより、それぞれのパラメータが独立となる確率変数 $\mathbf{f}'$ を定義できることが分かる。$\mathbf{f}'$ の標本が得られた後は、$\mathbf{f} = \mathbf{U}\mathbf{f}'$ により、もとの確率変数に戻せばよい。

以上の記述に基づいた、複数の確率変数についてランダムに標本を抽出する方法は以下のようになるであろう。

1. 共分散行列 $\mathbf{M}$ の対角化を行い、行列 $\mathbf{U}$ を得る。
2. $\mathbf{f}$ を線形変換した $\mathbf{f}'$ について、その平均値 $\bar{\mathbf{f}}'$ と共分散行列 $\mathbf{D}$ に基づいて標本を抽出する。ここで、$\mathbf{f}'$ はそれぞれが独立な確率変数であるので、それぞれ独立に標本抽出が可能である。
3. 得られた $\mathbf{f}'$ についての標本から、目的とする確率変数に関する標本を $\mathbf{f} = \mathbf{U}\mathbf{f}'$ の関係より求める。

なお、一般的には以下のような方法が採用されている。共分散行列の対角化の式を以下のように記述する。

$$\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{U}^T = \mathbf{U}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{U}^T = \mathbf{U}\mathbf{D}^{1/2}\left(\mathbf{D}^{1/2}\right)^T\mathbf{U}^T = \mathbf{U}\mathbf{D}^{1/2}\left(\mathbf{U}\mathbf{D}^{1/2}\right)^T = \mathbf{A}\mathbf{A}^T \tag{27}$$

ここで、$\mathbf{A}^{-1} = \mathbf{D}^{-1/2}\mathbf{U}^T$、$\left(\mathbf{A}^T\right)^{-1} = \mathbf{U}\mathbf{D}^{-1/2} = \left(\mathbf{A}^{-1}\right)^T$ なので、

$$\mathbf{A}^{-1}\mathbf{M}\left(\mathbf{A}^{-1}\right)^T = \mathbf{I} \tag{28}$$

が成り立つ。これより、$\mathbf{f}' = \mathbf{A}^{-1}\mathbf{f}$ なる変換を行った場合、$\mathbf{f}'$ の共分散行列は $\mathbf{I}$ となることが分かる。従って、標準正規分布に従う確率変数の標本を $\mathbf{f}'$ としたときに、それに対応する元の確率変数の標本は以下の式で求められる。

$$\mathbf{f} = \mathbf{A}\mathbf{f}' = \left(\mathbf{U}\mathbf{D}^{1/2}\right)\mathbf{f}' \tag{29}$$

この方法では、前述の方法のように線形変換後の確率変数の平均値 $\bar{\mathbf{f}}'$ を求めておく必要がないなどの利点がある。

問題 7：平均 5.38、標準偏差 0.55 の正規分布に従う 2 つの確率変数を考える。両者の相関を 0.8 として 100、1,000、10,000 個の標本を作成し、それら標本群から共分散行列を求め、標本数の増加に伴いもとの（母分布の）共分散行列の再現性が向上することを確認せよ [a] 。

[a] なお、標準偏差がいずれも σ であり相関がない 2 つの確率変数 x、y について、$x' = x$、$y' = Cx + \sqrt{1-C^2}y$ という変換を行って得られる x'、y' を考えると、これらは標準偏差が σ で相関が C となる。これについては実際に線形変換行列 $\mathbf{P}$ を

$$\mathbf{P} = \begin{pmatrix} 1 & 0 \\ C & \sqrt{1-C^2} \end{pmatrix} \tag{30}$$

として、x'、y' の共分散行列を計算し、そのような結果が得られること確認してみるとよい。問題 6 はこのようにしても解くことができるが、ここでは解説で述べられているように、もとの共分散行列の固有ベクトルから計算する方法を用いてもらいたい。

共分散行列とは、対角成分が分散、非対角成分が共分散に対応する行列である。共分散行列 $\mathbf{M}$ の要素 M_{ii} は f_i の分散になる。また f_i と f_j の相関係数 C_{ij} は共分散行列を用いて以下のように定義される。

$$C_{ij} = \frac{M_{ij}}{\sqrt{M_{ii}}\sqrt{M_{jj}}} = \frac{M_{ij}}{\sigma_i\sigma_j} \tag{31}$$

ここで σ_i は f_i の標準偏差を示す。

また共分散行列のような対称正方行列について特異値分解を行う場合、それはこの行列の固有値を求める操作と等価となる。すなわち、共分散行列 $\mathbf{M}$ の固有値を λ_i、固有関数を x_i とすると以下の方程式が成り立つ。

$$\mathbf{M}\mathbf{x}_i = \lambda_i\mathbf{x}_i \tag{32}$$

ここで、$\mathbf{X} = (\mathbf{x}_1\ \ \mathbf{x}_2\ \ \cdots\ \ \mathbf{x}_L)$、$\mathbf{D} = diag\,(\lambda_1, \lambda_2, \cdots, \lambda_L)$ とすると

$$\mathbf{M}\mathbf{X} = \mathbf{X}\mathbf{D} \tag{33}$$

が得られる。$\mathbf{X}^{-1} = \mathbf{X}^T$ より

$$\mathbf{X}^T\mathbf{M}\mathbf{X} = \mathbf{D} \tag{34}$$

となり、線形変換により対角成分がゼロの共分散行列が得られる。問題 7 では行列のサイズが小さいため固有値と固有ベクトルを解析的に得ることができることに留意すること。

最後に、確率変数の対数が正規分布に従う対数正規分布の場合について、標本の抽出方法を以下にまとめておく。ここで、確率変数 x の対数 $z = \ln(x)$ が正規分布に従うものとする。

1. x の平均値 μ_x と共分散 V_x から、z の平均値 μ_z と共分散 V_z を計算する。
2. z について、μ_z、V_z に基づいて標本の抽出を行う。
3. $x = \exp(z)$ より、z の標本に対応する x の標本を得る。

参考文献

[1] 東京大学教養学部統計学教室編、「基礎統計学 I、統計学入門」、東京大学出版会.

[2] 長家康展、「モンテカルロ計算の基礎理論及び実験解析への適用」、第 38 回炉物理夏期セミナーテキスト、(2006).

[3] 山本章夫、「不確かさ評価の基礎」、第 44 回炉物理夏期セミナーテキスト、(2012).

[4] 遠藤知弘、「ランダムサンプリング法で得られた不確かさの統計誤差」、私信.