コルモゴロフ-スミルノフ検定の数値実験

2020SS028 方田留騎亜
指導教員：小藤俊幸

1　はじめに

多くの統計的手法において, 正規分布に従うことを仮定されているため, データが正規分布に従うことを確認しておく必要がある. コルモゴロフ-スミルノフ検定によって統計的に正規乱数が正規分布に従うか考える. 最初の段階として一様分布を考え, その一様乱数からボックス・マラー法によって正規乱数を生成し, コルモゴロフ-スミルノフ検定によって, ボックス・マラー法による経験分布関数が標準正規分布に従うか考察する.

2　乱数の検定

以下，$x_1, x_2, \ldots, x_n$ は，線形合同法，Xorshift，メルセンヌ・ツイスターなどによって生成された (乱数列を変換して得られる) 区間 [0, 1) 上の一様乱数列を表すものとする.

2.1　経験分布関数

実数 x に対して,

$$F_n(x) \;=\; \Big(x \text{ 以下の } x_1, x_2, \ldots, x_n \text{ の個数}\Big)/n \qquad (1)$$

によって定まる関数を，**経験分布関数** (empirical distribution) という．乱数列を昇順にソートした数列をあらためて，$0 \le x_1 \le x_2 \le \cdots \le x_n < 1$ と書くことにすると，$F_n(x)$ は

$$F_n(x) \;=\; \begin{cases} 0 & (x < x_1) \\ 1/n & (x_1 \le x < x_2) \\ 2/n & (x_2 \le x < x_3) \\ \vdots & \vdots \\ (n-1)/n & (x_{n-1} \le x < x_n) \\ 1 & (x_n \le x) \end{cases} \qquad (2)$$

のように表される.

次は奥村の本 [1] にある線形合同法で生成した乱数列を, バブルソートを用いて, ソートするプログラムの経験分布関数である.

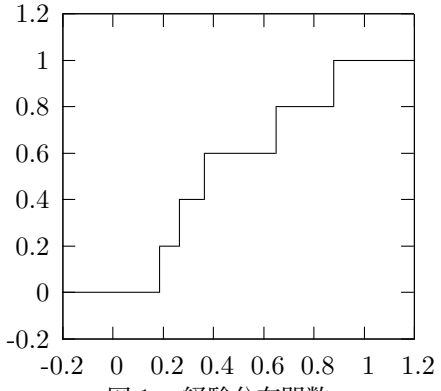

図 1　経験分布関数

上の図は sort23.c の計算結果から描いた経験分布関数のグラフである.

経験分布関数は, n が大きくなると，[0, 1) 上の一様分布の分布関数 $F(x) = x$ $(0 \le x < 1)$ に近づくと考えられる.

2.2　コルモゴロフ-スミルノフ検定

コルモゴロフ-スミルノフ検定は，標本値が連続な分布に従うかどうかを調べる一般的な検定法である．以下，区間 [0, 1) 上の一様分布に限定して説明する.

区間 [0, 1) 上の一様乱数の場合，経験分布関数と一様分布の分布関数 $F(x) = x$ $(0 \le x < 1)$ から

$$\begin{aligned} K_n^+ &= \sqrt{n}\ \sup_{0\le x<1}\Big(F_n(x) - x\Big), \\ K_n^- &= \sqrt{n}\ \sup_{0\le x<1}\Big(x - F_n(x)\Big) \end{aligned} \qquad (3)$$

のような量を求める.

乱数列 $x_1, x_2, \ldots, x_n$ が $(x_1 \le x_2 \le \ldots, \le x_n$ のように) ソートされているとすると，(3) は

$$\begin{aligned} K_n^+ &= \sqrt{n}\ \max_{1\le i\le n}\Big(\tfrac{i}{n} - x_i\Big), \\ K_n^- &= \sqrt{n}\ \max_{1\le x\le n}\Big(x_i - \tfrac{i-1}{n}\Big) \end{aligned} \qquad (4)$$

のように書き直される.

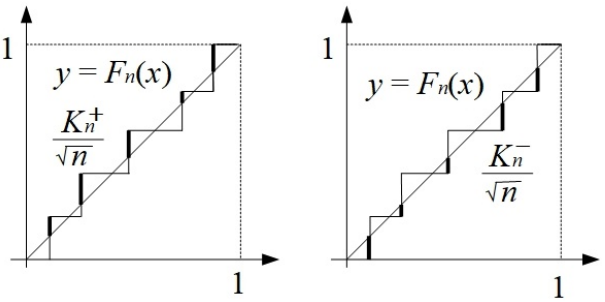

図 2　$K_n^+/\sqrt{n}$　　,　　$K_n^-/\sqrt{n}$

図形的に考えると，$K_n^+/\sqrt{n}$ は，$y = F_n(x)$ のグラフが $y = x$ の上側にはみ出した幅の最大値，$K_n^-/\sqrt{n}$ は，下側にはみ出した幅の最大値となる．K_n^+ と K_n^- がどのぐらい小さければ，$F_n(x)$ と $F(x) = x$ が近いと判断するのか，基準を以下のように決める.

(4) の K_n^+ と K_n^- の乱数 $x_1, x_2, \ldots, x_n$ を，[0, 1) 上の一様分布に従う独立な確率変数 $X_1, X_2, \ldots, X_n$ で置き換えたものを，同じ記号 K_n^+ と K_n^- で表すとき，実数 t に

ついて，

$$P\left(K_n^+ \le \frac{t}{\sqrt{n}}\right) = P\left(K_n^- \le \frac{t}{\sqrt{n}}\right) \\ = \frac{t}{n^n}\sum_{0\le k\le t} {}_n\mathrm{C}_k\,(k-t)^k\,(t+n-k)^{n-1-k} \tag{5}$$

が成り立つことが知られている ([2], p. 57).

$P\left(K_3^+ \le x\right) = 0.95$ となる点 x (95% 点) は $x = 1.1017$ である．

次頁の表に，線形合同法，Xorshift, メルセンヌ・ツイスターについて計算した K_{1000}^+, K_{1000}^- の値を示する．実数への変換は，いずれも関数 genrand_res53 を使用し，最初の 1000 個は飛ばして，1001 個めからの 1000 個の乱数で，(4) の値を計算した．$n = 1000$ の場合の，95% 点 (上側 5% 点）は，$x = 1.2239$ としてよいようなので ([2], p. 51)，いずれの方法も，この検定に合格したとみていいだろう．

表 1　計算結果

方法	線形合同法	Xorshift	MT
K_{1000}^+	0.894447	0.596785	0.160132
K_{1000}^-	0.410979	0.726386	1.010042

3　正規乱数の検定

3.1　ボックス・マラー法

正規乱数は一様乱数を変換して生成するのが一般的である. この変換方法として, ボックス・マラー法と極性マーセイリア法がよく知られている [4] . 今回はメルセンヌ・ツイスターを用いて一様乱数を生成し, ボックス・マラー法を用いて正規乱数を生成する. ボックス・マラー法のアルゴリズムは (6) 式のようになる.

u_j, v_j を区間 $[0, 1)$ 上の一様乱数（ただし，$u_j > 0$) とし，

$$\begin{cases} x_j = \sqrt{-2\log u_j}\,\cos(2\pi v_j) \\ y_j = \sqrt{-2\log u_j}\,\sin(2\pi v_j) \end{cases} \tag{6}$$

とすれば，x_j, y_j は標準正規分布に従う（独立とみなせる）乱数となる．この正規分布に従う確率変数を生成させる方法を**ボックス・マラー法**と呼ぶ.

ボックス・マラー法による正規乱数生成のプログラム

```
double rand_gauss()
{
    static int iset = 0;
    static double nod1;
    double u1, u2;
    iset = 1 - iset;
    if (iset == 0) return nod1;
    u1 = sqrt( - 2.0 * log(genrand_res53()));
    u2 = 2.0 * Pi * genrand_res53();
    nod1 = u1 * sin(u2);
    return u1 * cos(u2);
}
```

3.2　検定結果

標準正規分布の分布関数とボックス・マラー法による正規乱数の経験分布関数から K_n^+, K_n^- を求める.

図として表すと,$n = 1000$ のとき図 3 のようになり,$n = 100$ のとき図 4 のようになる.

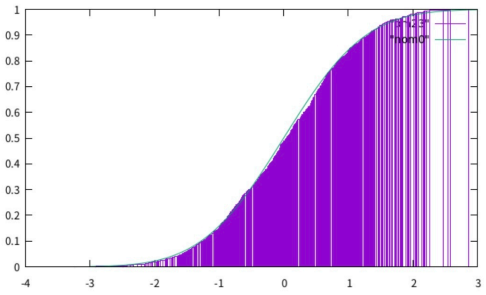

図 3　分布関数 ($n = 1000$)

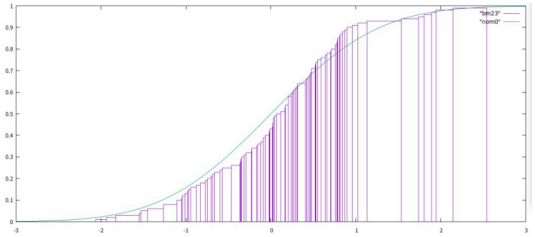

図 4　分布関数 ($n = 100$)

$n = 100$ のとき,$n = 1000$ のときそれぞれの K_n^+, K_n^- を求めると以下のようになる.

$$K_{100}^+ = 0.852194, \quad K_{100}^- = 0.967494$$
$$K_{1000}^+ = 0.462714, \quad K_{1000}^- = 1.106753$$

Knuth の本 [2] より,$n = 30$ の場合の 95% 点は,$x = 1.1916$ とし,$n = 1000$ の場合の 95% 点は,$x = 1.2239$ としてよいことから,$n = 100$ のとき,$n = 1000$ のときともに,コルモゴロフ-スミルノフ検定に合格したといえる.

4　おわりに

本研究でコルモゴロフ-スミルノフ検定の方法が分かり, プログラム自体も直線からはみ出した幅の最大値を取るようなものであることがわかった. また, コルモゴロフ-スミルノフ検定によって, ボックス・マラー法による経験分布関数が標準正規分布に従うか考察し,$n = 100$ のときも $n = 1000$ のときも正規分布に従うことが分かった.

参考文献

[1] 奥村晴彦：『【改訂新版】C 言語による標準アルゴリズム事典』(第 2 版)，技術評論社，2018.

[2] D. E. Knuth (渋谷政昭 訳)：『The Art of Computer Programming, 第 3 分冊，準数値算法/乱数』，サイエンス社，1981.

[3] Ya.G. シナイ (森 真 訳)：『シナイ確率論入門コース』，シュプリンガー・フェアラーク東京，1995.

[4] 三井斌友・小藤俊幸・齊藤善弘：『微分方程式による計算科学入門』. 共立出版, 2004 年.