

データの統計処理（応用）

重田 出

学習の目標

頻度分布，平均・標準偏差，最小二乗法，データの移動平均を学習し，理解する。

3 頻度分布

プログラム 3 で，1 から 6 の発生回数を数えるプログラムを考えてみる。もっとも単純な方法は，各数値に対応したカウンタを用意し，乱数を発生させるごとにその数値を調べ，その数値に対応したカウンタの数値を 1 増やす，という方法である。プログラム 1 では，カウンタは 1 が発生したときに 1 増やすもの，2 が発生したとき 1 増やすもの， $\dots$ ，6 が発生したときに 1 増やすもの，まで 6 つ用意すればよい。数字の判定には，if 文を用いればよい（複数の数を判定するときには，switch 文 + case 文を使う方法もある）。

演習 5 プログラム 3 で 100 回乱数を発生させ，頻度分布を求めるプログラムを作成せよ。

配列を使って頻度分布を作成するプログラム例をプログラム 3 に示しておく。配列を使うとき注意することは，配列の範囲が「0」から「配列の最大値 - 1」であることである。この範囲をはずれた位置に書き込もうとすると，プログラムは異常終了する。

プログラム 4 配列を用いた頻度分布作成プログラム

```
#include<stdio.h>
#include<stdlib.h>          // rand を使うときのヘッダ

int main()
{
    int sai, i, j;           // 整数変数の定義
    int hist[6];            // 頻度分布計数用の配列

    for (sai=1; sai<=6; sai++) hist[sai-1]=0; // 配列の初期化

    for (i=0; i<100; i++)
    {
        j=rand();           // 乱数を発生させる
        sai=j%6+1;          // 1 から 6 までの数にする
    }
}
```

```

        hist[sai-1]++;          // sai の値に対応する配列を+1
        printf("出た目は%d\n", sai);          // sai を表示
    }

    for (sai=1; sai<=6; sai++)
        // 結果の表示
        printf("出た目が%d の数は%d\n", sai, hist[sai-1]);

    return 0;
}

```

4 平均値と標準偏差

データが N 個あり、各データの値が x_i ($i = 1, 2, 3, \dots, N$) であるとする、平均値 $\bar{x}$ は、

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (1)$$

により求められる。

正規分布では、データのばらつき具合を表す量として、標準偏差が用いられる。標準偏差 σ は、

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2 - \bar{x}^2} \quad (2)$$

により与えられる。 $(x_i - \bar{x})^2$ はデータの離れ具合（距離）と考えることができるので、 $\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$ は、全体としてデータが平均からどの程度ばらついているか（平均距離）を意味している。

演習 6 プログラム 4 で乱数を発生させ、平均値を求めるプログラムを作成せよ。次に、繰り返し回数を 50 回、100 回と変えて、値がどうなるか調べよ。

（ヒント）繰り返し回数が 50 回のときの平均の求め方は、

```

mean=0.0;
for (i=0; i<50; i++) mean+=(double)sai;
mean=mean/(double)50;

```

なお、これは一例であり、別の書き方もある。繰り返し回数が 100 回のときも同じプログラムが使えるように、回数の変更を容易にする工夫をせよ。`math.h` のインクルードを忘れないこと（やり方は後述のプログラム 5 を参照）。

また、`mean+=(double)sai;` の部分を `mean+= (double)sai;` とすると、うまく動かない。なぜか？

偏差値とは

偏差値 (Standard score または T-score) とは, ある数値が母集団の中でどれくらいの位置にいるかを表した無次元数のことである。また, 平均値が 50, 標準偏差が 10 となるように標本変数を規格化したものである。

偏差値 T_i は次の式で求まる。

$$T_i = \frac{10(x_i - \bar{x})}{\sigma} + 50. \quad (3)$$

ただし,

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i, \quad (4)$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}. \quad (5)$$

ここで, N : 母集団の大きさ, x_i : 個々の値, $\bar{x}$: 母平均, σ : 母標準偏差である。なお, 分母 σ は標準偏差であり, 分子は偏差である。偏差が 0 であるとき (ある標本値が平均値に等しいとき) は偏差値は 50 となる。また, 標準偏差が 0 である場合, つまり全ての標本値が同じ値であるときはこの式では偏差値は定義できないが, その場合は便宜上全ての標本値の偏差値を 50 と考えることがある。

5 正規乱数の発生

ある長さのものをメジャーで繰り返し測った時, 真値に近いほどその発生頻度は高くなっている。このような繰り返し測定を行ったときの誤差の分布は, 正規乱数の特性を持っていることが多い。このため, 正規確率分布を持つ乱数が必要になる場合がある。しかし, C 言語にあらかじめ用意されている乱数を発生させる関数は, 一様乱数 `rand` のみである。正規乱数は, $(0, 1]$ の範囲の 2 組の一様乱数 α と β に対して,

$$\sqrt{-2 \cdot \ln \alpha} \cdot \sin(2\pi\beta) \quad (6)$$

$$\sqrt{-2 \cdot \ln \alpha} \cdot \cos(2\pi\beta) \quad (7)$$

という変換 (ボックス・ミュラー法/Box-Muller transform) を施すことにより求めることができる。この変換では, 平均が 0 で標準偏差が 1 である互いに独立な 2 つの正規乱数を得られる。ここで, $\ln$ は自然対数をとることを表している。

演習 7 一様乱数から正規乱数を生成するプログラムを作成せよ。

(ヒント) `rand()` で発生させた一様乱数は、一つ飛びに値を使ってもやはり一様乱数である。したがって、 α と β は同じ `rand()` から求めても良い。正規乱数を 1 組発生させるときは 2 つの変換のうち、どちらを使ってもよい。 $(0, 1]$ の範囲の一様乱数 a は、

```
a=(double)(rand()+1)/(double)(RAND_MAX+1);
```

として得られる。また、 $\sqrt{x}$ は `sqrt(x)`、自然対数 $\ln(x)$ は `log(x)`、正弦関数は `sin(x)` である。ここで、 x は倍精度実数、正弦関数のときの x の単位はラジアンである。

プログラム 5 正規乱数を生成するプログラム例

```
#include<stdio.h>
#include<stdlib.h>          // rand を使うときのヘッダ
#include<math.h>            // 数学関数を使うときのヘッダ

int main()
{
    int i;                  // 整数変数の定義
    double x, a, b;

    for (i=0; i<100; i++)
    {
        a=(double)(rand()+1)/(double)(RAND_MAX+1);
        b=(double)(rand()+1)/(double)(RAND_MAX+1);
        x=sqrt(-2.0*log(a))*cos(2.0*3.141593*b);
        printf("a,b,x=%lf,%lf,%lf\n",a,b,x); // 結果を表示
    }

    return 0;
}
```

演習 8 正規乱数の性質を持つ数列を生成し、頻度分布を求めるプログラムを作成せよ。また、平均 $\pm$ 標準偏差の範囲に何割のデータが入っているか求めるプログラムを作成せよ。

(ヒント) 演習 7 で作成したプログラムでは、平均 0、標準偏差 1 の正規乱数となっている。一様乱数とは異なり、正規乱数では出現する値の上限、下限がないため、非常に大きな値、小さな値が現れることがある。頻度分布を求めるときは、範囲外の値に対する処理を考えておく必要がある。

6 データの直線近似

図 5 のような $x-y$ の二次元座標上に、 N 個の点があったとし、 x と y の関係を傾き a 、切片 b の直線で近似するとする。このとき、近似直線から各点がどれだけ

離れているかは、

$$E^2 = \{y(x_i) - y_i\}^2 = \{(a \cdot x_i + b) - y_i\}^2 \quad (8)$$

という量で評価することができる。

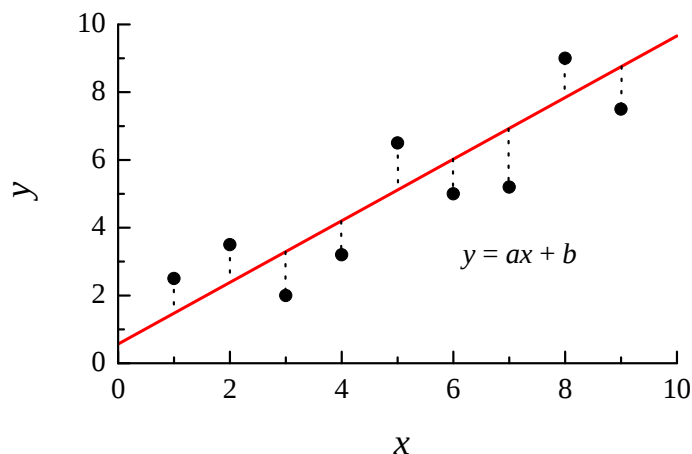


図 5: 2次元のデータとそれを近似する直線。

演習 9 下記のデータを、横軸 x 、縦軸 y としてグラフ用紙に書き込み、この点を近似する直線を適当に引く。このときに、図 5 のようにして直線と各点との距離をグラフから読み取り、 E^2 の値を求める。次に、切片 b を一定値（ここでは、0.57）として、傾き a を変えた直線を引き、 E^2 の値を求める。この操作を様々な傾き a に対して行い、 E^2 が a に対してどのように変化しているか調べよ。

（コメント） 1 人でこの作業を行うのは大変なので、切片 b を同じにしたままクラスの各自で異なる傾きに対する E^2 を分担して計算する。

x	1	2	3	4	5	6	7	8	9	10
y	2.0	2.7	2.4	3.5	5.8	6.0	7.2	7.4	8.8	9.9

最適値（傾き）= 0.91,（切片）= 0.57

演習 10 同じデータに対して、傾き a を一定値（ここでは、0.91）として、切片 b を変えた直線を引き、 E^2 の値を求める。この操作を様々な切片 b に対して行い、 E^2 が b に対してどのように変化しているか調べよ。

演習 11 以上の結果から、傾き a および切片 b の最適値を求めるにはどうしたら良いか考えよ。

演習 12 以上の考察に基づき、データを最もよく近似する直線を求めるプログラムを作成せよ。

7 移動平均 (Running Mean)

図 6(a) に示すような時系列データ x_i ($i = 1, 2, 3, \dots$) を考える。この図を見ると、大まかには正弦波状の変化をしているが、それ以外にも細かくでこぼこした変動が見られる。

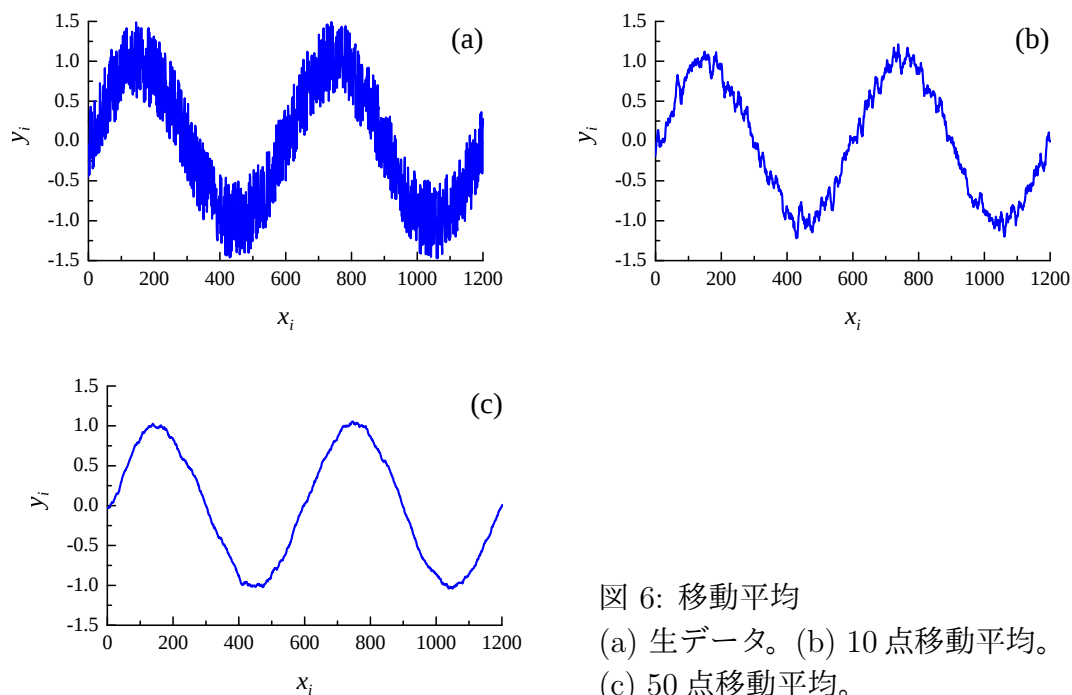


図 6: 移動平均

(a) 生データ。(b) 10 点移動平均。

(c) 50 点移動平均。

この時系列データに対し、それぞれの時刻において前後 N 点の平均をとり、その結果をそれぞれの時刻のデータとする操作

$$y_i = \frac{1}{2N+1} \sum_{i=-N}^N x_i \quad (9)$$

を行うと、細かくでこぼこした変動が小さくなり、ゆっくりとした変動のみが見えるようになる (図 6(b),(c))。このような操作を「移動平均」と呼んでいる。移動平均は、低い周波数成分だけを通す、すなわち低域通過フィルタ (ローパス・フィルタ) の役割をしている。