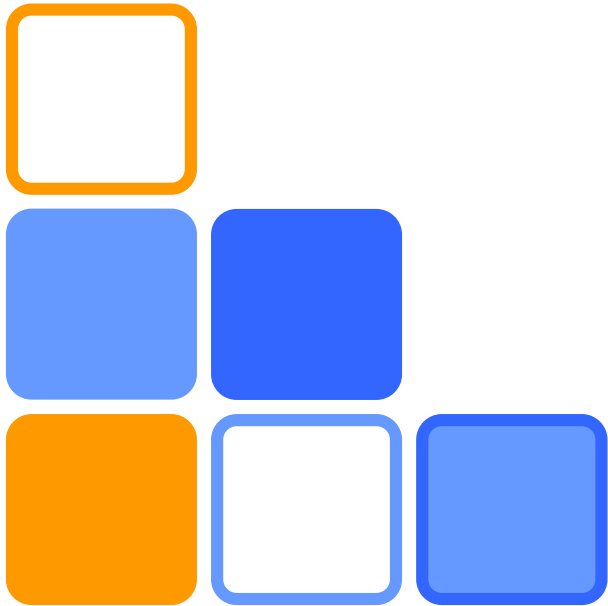


薬学情報処理演習 第2回

表計算ソフトによる統計 処理



奥菌 透

コロイド・高分子物性学

データの整理

参考文献: 東京大学教養学部統計学教室編
「統計学入門」(東京大学出版会, 1991)

□ 平均値と分散

- N 個の(数値)データ $x_1, x_2, \dots, x_N$
- 平均値: $\bar{x} = \frac{x_1 + x_2 + \dots + x_N}{N}$
- 分散: $\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_N - \bar{x})^2}{N}$ (σ : 標準偏差)

□ 度数分布

- 値 $X_1, X_2, \dots, X_n$ をとる度数(頻度): $F_1, F_2, \dots, F_n$
- $\bar{x} = \frac{F_1 X_1 + F_2 X_2 + \dots + F_n X_n}{F_1 + F_2 + \dots + F_n}$ $N = F_1 + F_2 + \dots + F_n$
- $\sigma^2 = \frac{F_1 (X_1 - \bar{x})^2 + F_2 (X_2 - \bar{x})^2 + \dots + F_n (X_n - \bar{x})^2}{F_1 + F_2 + \dots + F_n}$

代表値 X_i	度数 F_i
1	1
2	4
3	15
4	9
5	1

□ チェビシェフの定理

$|x_i - \bar{x}| \leq \lambda \sigma$ ($\lambda > 1$) を満たすデータの個数は
 $N(1 - 1/\lambda^2)$ よりも大きい。

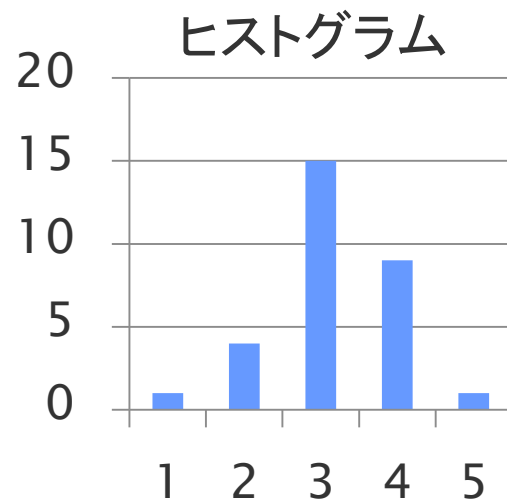
データの分布

離散的なデータ

- 度数分布 F_i
- 確率 $f_i = F_i/N$

$$\bar{x} = \sum_{i=1}^n X_i f_i \quad \sigma^2 = \sum_{i=1}^n (X_i - \bar{x})^2 f_i$$

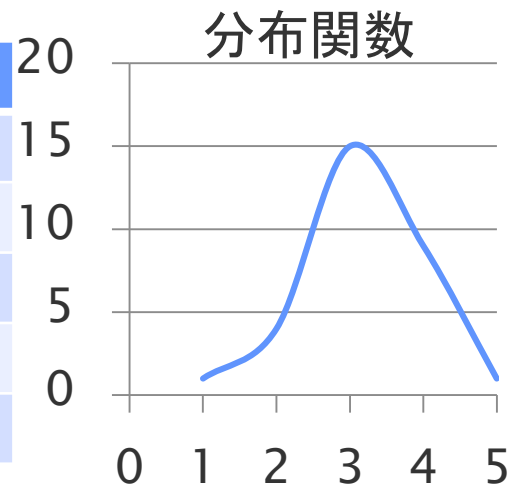
代表値 X_i	度数 F_i
1	1
2	4
3	15
4	9
5	1



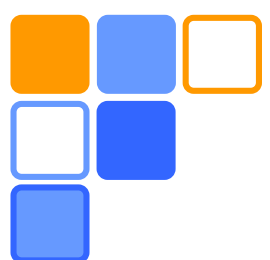
連続的なデータ

- 度数分布 → 分布関数 $F(x)$
($N \rightarrow \infty, \Delta x = X_{i+1} - X_i \rightarrow 0$)
- 確率密度 $f(x) = F(x)/\tilde{N}$
($\tilde{N} = \int F(x)dx$)

級	X_i	F_i
$0 < x \leq 1$	0.5	1
$1 < x \leq 2$	1.5	4
$2 < x \leq 3$	2.5	15
$3 < x \leq 4$	3.5	9
$4 < x \leq 5$	4.5	1



$$\bar{x} = \int x f(x) dx \quad \sigma^2 = \int (x - \bar{x})^2 f(x) dx$$



疑似乱数の生成

- コンピュータ上でランダムな数(乱数)を次々に生成し、ランダムなデータを作ることができる。エクセルでは RAND() という関数を用意されている。
- RAND()で生成される乱数は一様分布関数

$$f(x) = 1 \quad (0 \leq x < 1)$$

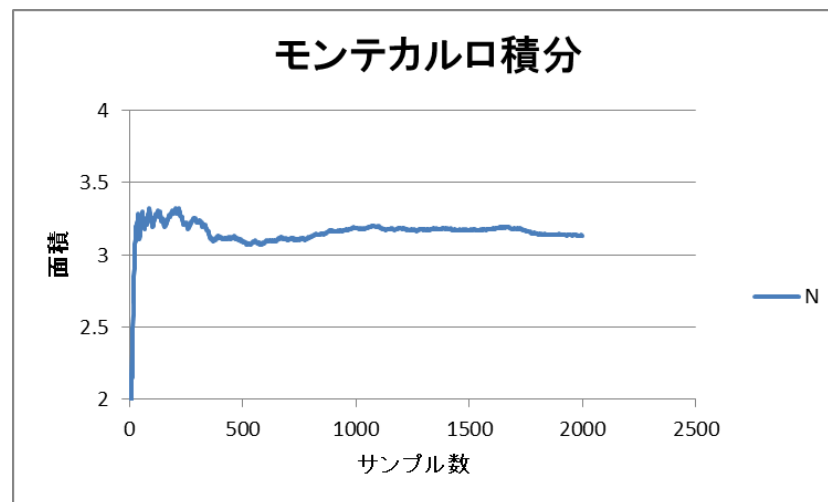
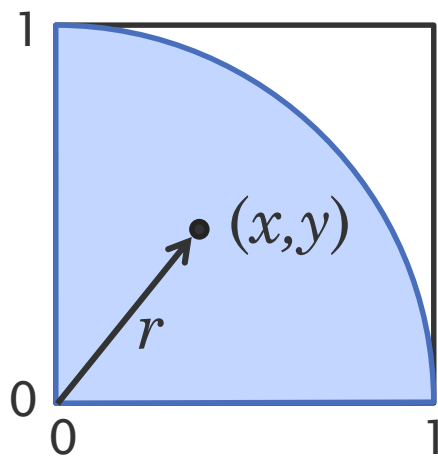
に従い、平均と分散は、

$$\bar{x} = \int_0^1 x f(x) dx = \frac{1}{2} \quad \sigma^2 = \int_0^1 (x - \bar{x})^2 f(x) dx = \frac{1}{12}$$

となるので、平均0の一様乱数(RAND()-0.5)を12個足し合わせたものは、近似的に、平均0分散1の正規分布に従う乱数(正規乱数)となっている。

演習課題1

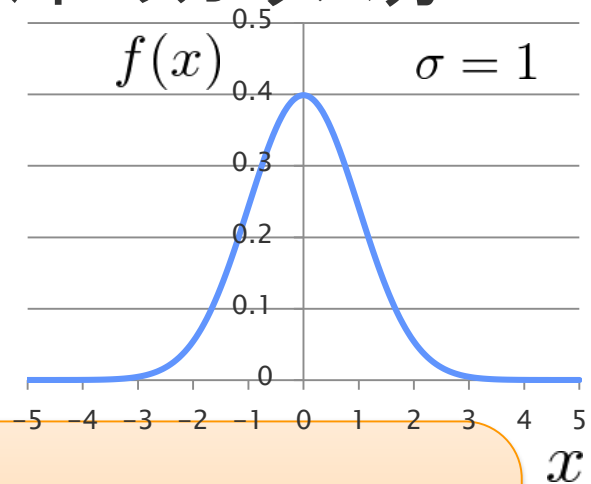
- 一様乱数の性質を利用して半径1の円の面積を求めてみよう(モンテカルロ積分)。
- 2つの乱数 (x, y) を発生させそれを xy 平面上の点とする。ただし、 $0 \leq x < 1$, $0 \leq y < 1$ とする。
- $r = \sqrt{x^2 + y^2}$ が $r \leq 1$ となる点の個数を n 、点の総数 N とし、 $S_N = 4n/N$ として円の面積を求める。
- N を大きくすると S_N は π に近づくことを確かめよ。



正規分布(ガウス分布)

- 実験的に測定される量には“ばらつき”がある。
ばらつき＝平均値からのずれは以下のガウス分布に従うことが多い。なぜか？

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \quad (\sigma^2 : \text{分散})$$



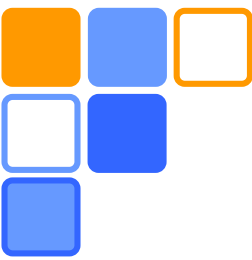
中心極限定理

n 個の独立な確率変数 u_i (分散 s_i^2 平均値 0) からなる確率変数

$$x_n = (u_1 + u_2 + \cdots + u_n) / \sqrt{\sigma_n^2} \quad \sigma_n^2 = s_1^2 + s_2^2 + \cdots + s_n^2$$

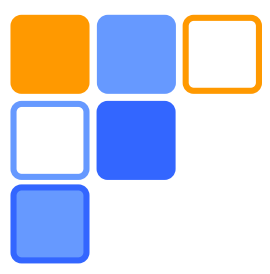
は、 $n \rightarrow \infty$ で分散 1, 平均値 0 の正規分布に従う。

- ばらつき＝多数の確率的事象の和



演習課題2

- 関数RAND()を用いて平均0の一樣分布に従う乱数データを作成する。
- 上記のデータから正規分布をもつ乱数データを作成する。
- 上記のデータに対する分布関数(正規分布と余裕があれば一樣分布も)を描く。
 - 分析ツール(後述)を用いて、度数分布表を作る。
 - 規格化された分布関数のデータを計算する。
 - 得られた分布関数のデータをグラフに描き、理論曲線と比較する。
 - (余裕があれば)平均値と分散を計算し、理論値と比較する。

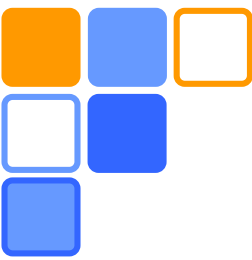


データの作成

- RAND()を使って平均0の一樣乱数を生成する。
 - 1列数千個のデータを12列作る。
- 12個の一樣乱数を足し合わせて正規乱数を生成する。

Uniform 1	Uniform 2	...	Uniform 12	Normal
=RAND()-0.5	=RAND()-0.5	...	=RAND()-0.5	=SUM(A2:L2)
...	...	...	...	...





度数分布表の作成

- エクセルで度数分布を作る方法はいろいろあるが、ここでは「分析ツール」を使う。これを使用可能とするには、**ファイル/オプション/アドイン/設定** で「分析ツール」を選択し「OK」をクリックする。
- データ区間(級)を作成する。
- 分析ツールを使う
 - **データ/分析/データ分析/ヒストグラム**
 - 入力範囲、データ区間を指定
 - 出力先を選択・指定

データ区間
-4.5
-4.3
-4.1
-3.9
...
4.3
4.5

分布関数データの作成

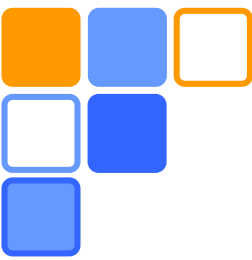
- 分析ツールで得られた度数分布表を用いる。
- 代表値、規格化された分布関数データを作成

規格化: $\sum_n F_n \Delta x = \mathcal{N} \quad \Rightarrow \quad \sum_n f_n \Delta x = 1, \quad f_n = F_n / \mathcal{N}$

データ区間	頻度	代表値 x	分布 f(x)
-4.5	0		
-4.3	1	=(A2+A3)/2	=B2/\$B\$49
...	...	...	...
4.5	0	...	...
次の級	0		
積分値	=SUM(B2:B47)*0.2		

$$f_n = F_n / \mathcal{N}$$

↑ $\mathcal{N}$



理論曲線データの作成

- 一様分布と正規分布のデータを作成する。

代表値 x	一様分布
-4.5	=IF(AND(C3>-0.5, C3<0.5), 1, 0)
...	...

IF(条件, 値1, 値2) : 条件が真であれば値1を, 偽であれば値2を返す。
AND(条件1, 条件2) : 条件1が真かつ条件2が真であれば真の値を返す。
したがって上の例は、セルC3の値が-0.5より大きく、0.5より小さいならば、セルに1が入力され、それ以外では0が入力される。

代表値 x	正規分布
-4.5	=NORM.DIST(C3, 0, 1, FALSE)
...	...

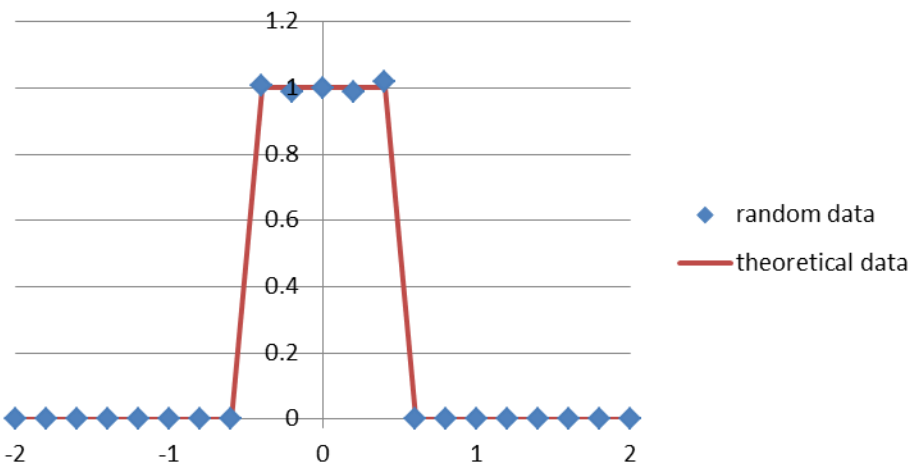
関数NORM.DIST($x, \bar{x}, \sigma^2$, FALSE) は x の値に対する平均 $\bar{x}$ 分散 σ^2 の正規分布 $f(x)$ の値を返す。

グラフの作成

分布関数のグラフ

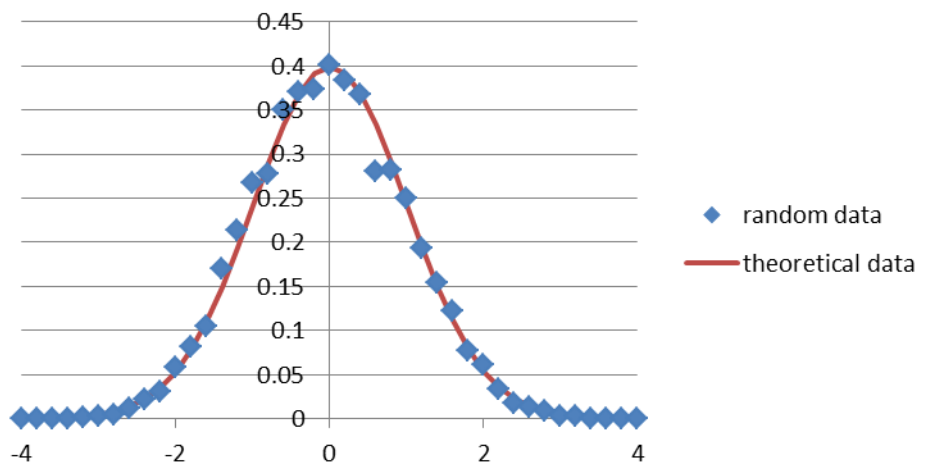
- 横軸に代表値、縦軸に確率密度分布をとる。
- 理論値と度数分布から得られたデータを比較する。

uniform distribution



平均値: 0.00085
分散: 0.08058 (分布関数から計算した値)
データ数: 60000

normal distribution



平均値: -0.01044
分散: 1.002027 (分布関数から計算した値)
データ数: 5000