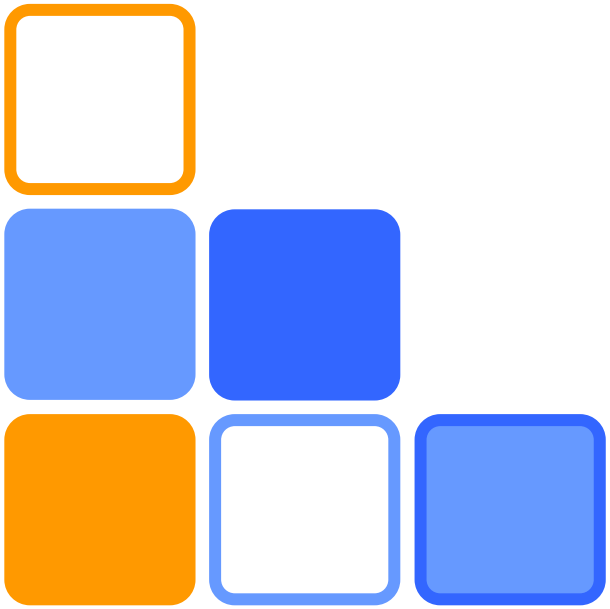
薬学情報処理演習　第2回

表計算ソフトによる統計処理



奥薗　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 $x_1, x_2, \cdots, x_N$
  - 平均値： $\bar{x} = \frac{x_1+x_2+\cdots+x_N}{N}$
  - 分散： $\sigma^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_N-\bar{x})^2}{N}$ （σ：標準偏差）
- 度数分布
  - 値$X_1, X_2, \cdots, X_n$ をとる度数（頻度）： $F_1, F_2, \cdots, F_n$
  - $\bar{x} = \frac{F_1X_1+F_2X_2+\cdots+F_nX_n}{F_1+F_2+\cdots+F_n}$ $\quad N = F_1 + F_2 + \cdots + F_n$
  - $\sigma^2 = \frac{F_1(X_1-\bar{x})^2+F_2(X_2-\bar{x})^2+\cdots+F_n(X_N-\bar{x})^2}{F_1+F_2+\cdots+F_n}$
- チェビシェフの定理

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

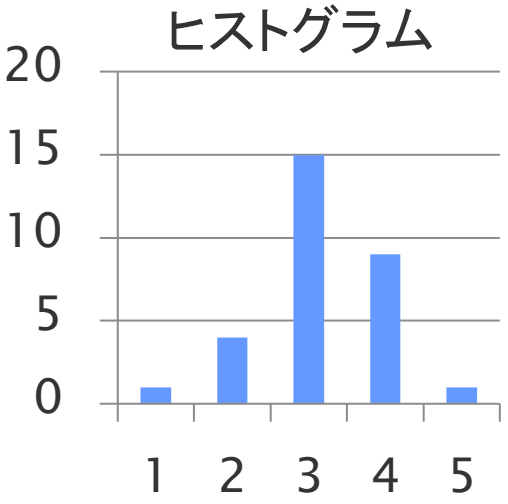
代表値 X_i	度数 F_i
1	1
2	4
3	15
4	9
5	1

データの分布

- 離散的なデータ
  - 度数分布 F_i
  - 確率 $f_i = F_i/N$

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

代表値 X_i	度数 F_i
1	1
2	4
3	15
4	9
5	1

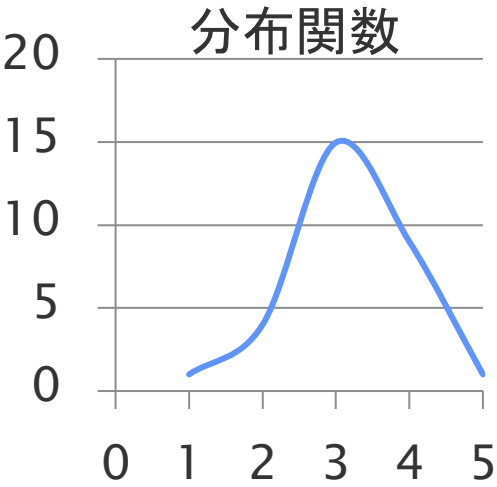


- 連続的なデータ
  - 度数分布→分布関数 $F(x)$

    $(N \to \infty, \Delta x = X_{i+1} - X_i \to 0)$
  - 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

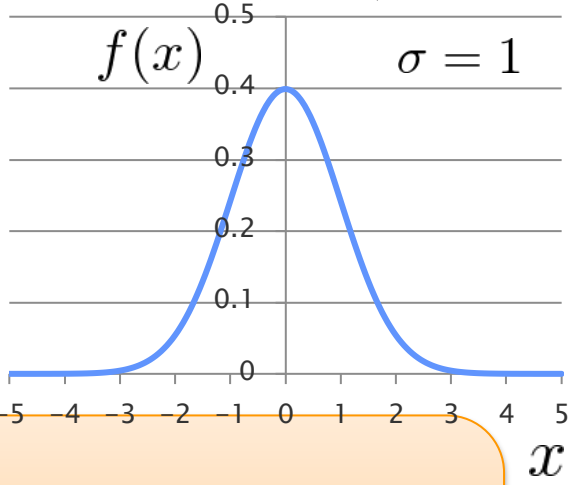


$$\bar{x} = \int x f(x)dx \qquad \sigma^2 = \int (x - \bar{x})^2 f(x)dx$$

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？

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



- 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

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

は、$n \to \infty$ 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

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() という関数が用意されている。
- RAND()で生成される乱数は一様分布関数

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

に従い、平均と分散は、

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

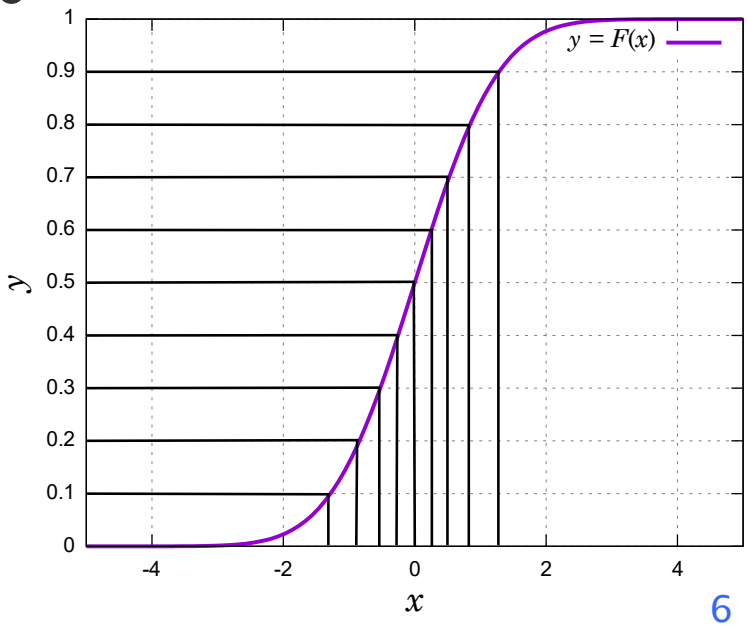
となるので、平均0の一様乱数（RAND()−0.5）を12個足し合わせたものは、近似的に、平均0分散1の正規分布に従う乱数（正規乱数）となっている。

Excel での正規乱数発生方法

- データ分析ツールの利用（関数ではない）
  - *データ/分析/データ分析/乱数発生*
- 中心極限定理の応用
  - （12個の一様乱数の和）-6
  - =rand()+rand()+…+rand()-6.0
- 逆関数法
  - =norm.inv(rand(),0,1)または
    =norm.s.inv(rand())を用いる
  - 分布$f(x)$の累積分布を$F(x)$とする

$$y = F(x) = \int_{-\infty}^{x} f(u)du$$

$$\frac{dy}{dx} = f(x),\ dy = f(x)dx$$



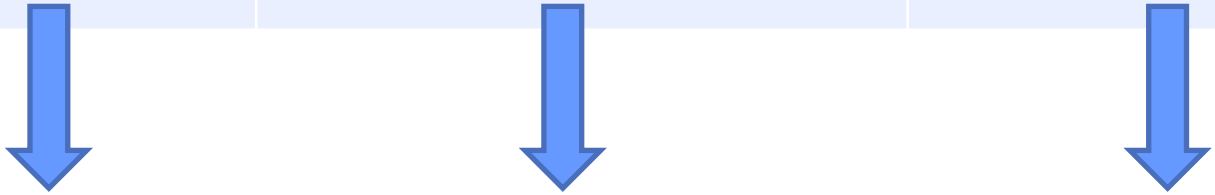
演習課題

- 上記の3つの方法で、それぞれN個の正規乱数を発生させる（$N = 10000$ 程度）。
- 上で得られたデータに対する分布関数のグラフを描く。
  - 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

- 分析ツールを使って正規乱数を発生
- RAND()を使って12個の一様乱数を足し合わせて正規乱数を生成
- 逆関数法により正規乱数を発生

excel	rand	inv.func
（分析ツール）	=rand()+…+rand()-6.0	=norm.s.inv(rand())
…	…	…



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

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		

$f_n = F_n / \mathcal{N}$

$\mathcal{N}$

Δx：データ区間の増分値

理論曲線データの作成

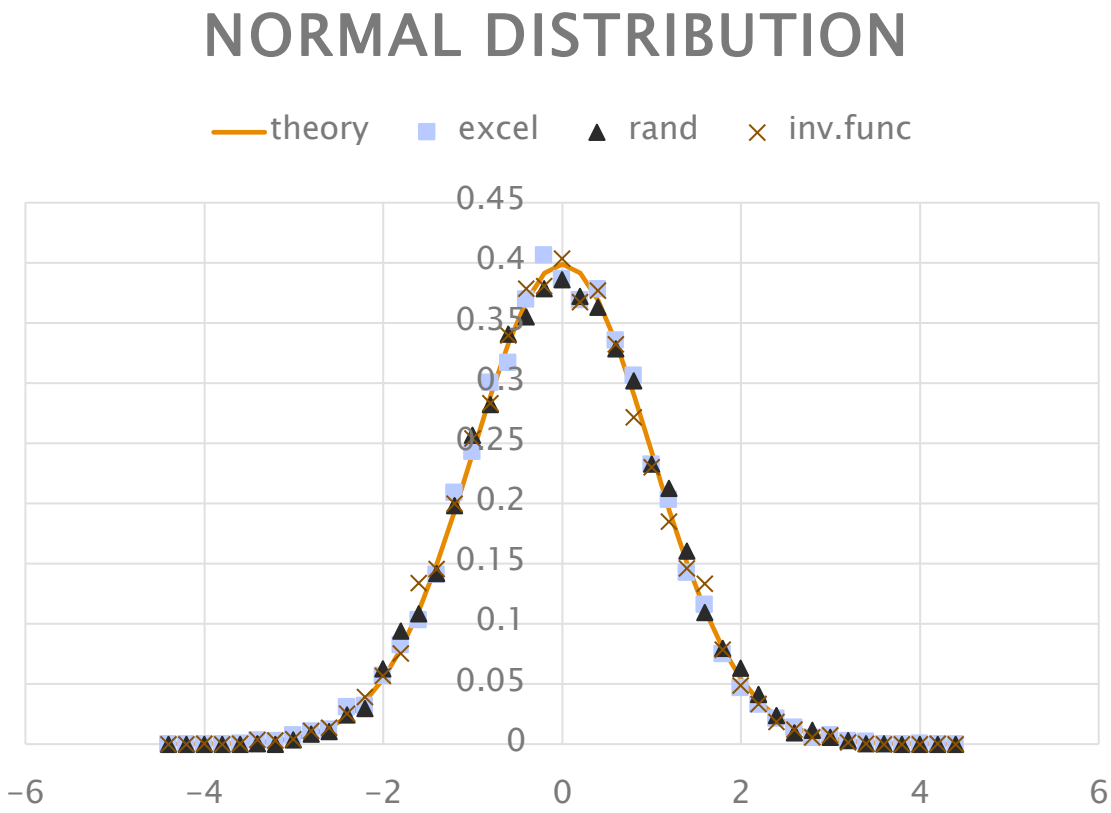
- 正規分布のデータを作成する。

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。
  - NORM.S.DIST(x, FALSE) は NORM.DIST($x, 0, 1$, FALSE) と同じ。
  - 最後の引数がFALSEの場合、(密度)分布関数を、TRUEの場合、累積分布関数を返す。

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。



方法	平均	分散
theory	0	1
excel	-0.00982	1.007028
rand	0.01078	1.019344
inv.func	-0.02518	1.015754

データ数：$N = 10000$

分析ツールを使わずに分布関数のデータを作成する。(補足)

- 元データの作成(データの作成参照)
- COUNTIF(範囲,検索条件)を使って累積度数分布を作成。規格化して累積分布関数を得る。
- 累積分布関数を微分して確率密度関数を得る。

	B	C	D	E	F	G
1	正規乱数	データ区間	累積度数	累積分布	代表値	確率密度
2	-1.52061	-4.5	0	0		
3	-1.37599	-4.3	0	0	-4.4	0
4	2.411254	-4.1	0	0	-4.2	0
5	-0.64805	-3.9	0	0	-4	0
6	0.113184	-3.7	0	0	-3.8	0
7	0.397722	-3.5	2	0.0002	-3.6	0.001
8	0.260922	-3.3	6	0.0006	-3.4	0.002

[D2]=COUNTIF($B$2:$B$10001,"<="&C2)
B列のデータのうちC2の値以下であるデータの数を返す。
[E2]=D2/D$47
累積度数の最後の値(=全データ数)で割って規格化する。
[G3]=(E3-E2)/0.2
累積分布を微分する。0.2はデータ区間の増分値。