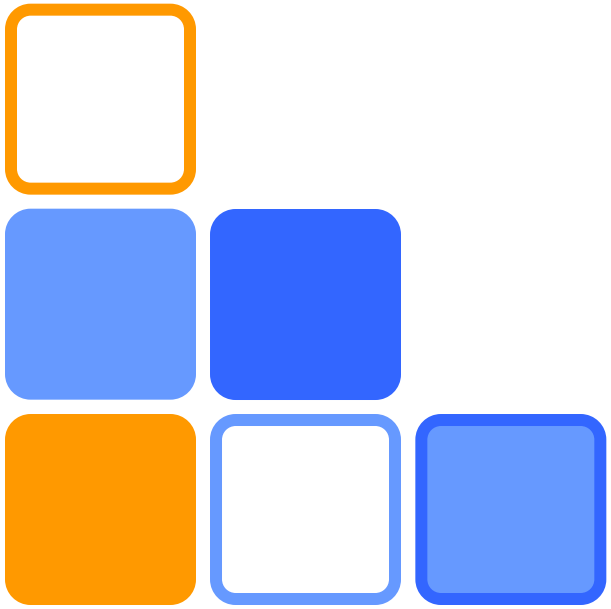
薬学情報処理演習　第2回

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

基本的な統計量

- N 個のデータ: $x_1,\ x_2, \ldots, x_N$

平均値: $\overline{x} = \dfrac{x_1 + x_2 + \cdots + x_N}{N}$

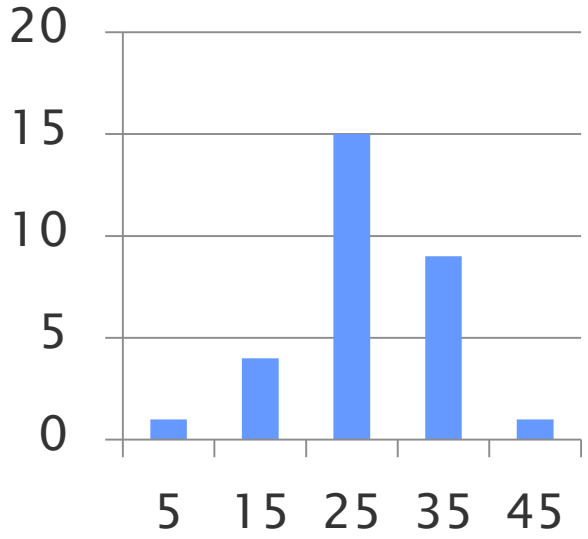
分散: $\sigma^2 = \dfrac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2}{N}$ （σ :標準偏差）

ただし、分散の推定値を計算するときは、分母の N を $N-1$ にする。

- 度数分布

級（class）	度数（frequency）
$0 < x_i \leq 10$	1
$10 < x_i \leq 20$	4
$20 < x_i \leq 30$	15
$30 < x_i \leq 40$	9
$40 < x_i \leq 50$	1

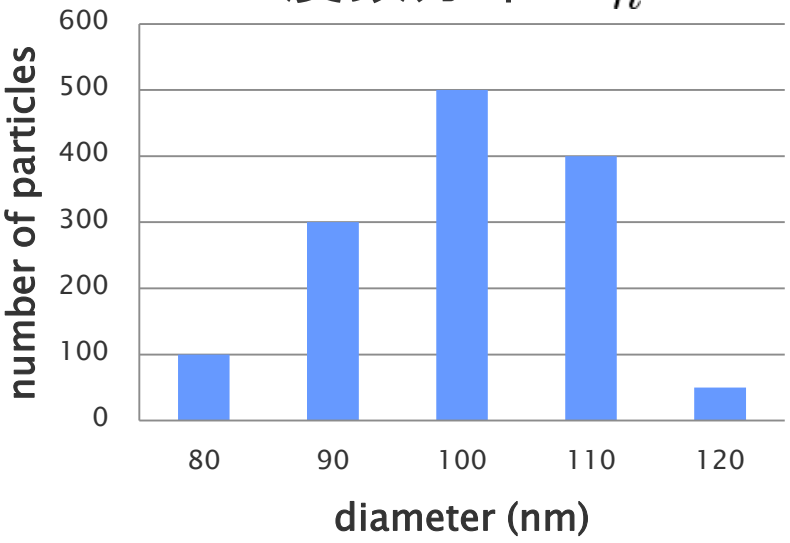
ヒストグラム



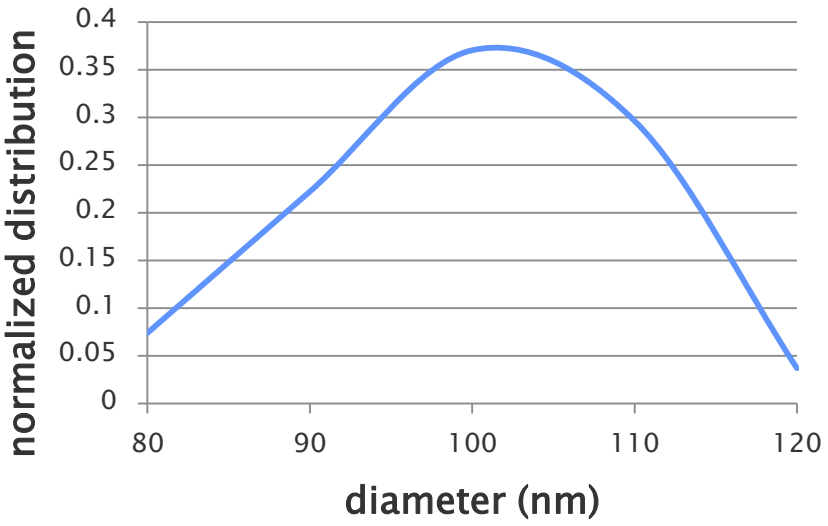
連続的な値を取り得るデータに対する分布関数

- 例：多数のコロイド粒子の粒径 x_i の測定値

度数分布 F_n



分布関数（確率密度）$f(x)$



F_n ：粒径が $x_n < x_i \leq x_n + \Delta x$ を満たす粒子の個数

$\Delta x = x_{n+1} - x_n$

$N \to \infty$

$\Delta x \to 0$

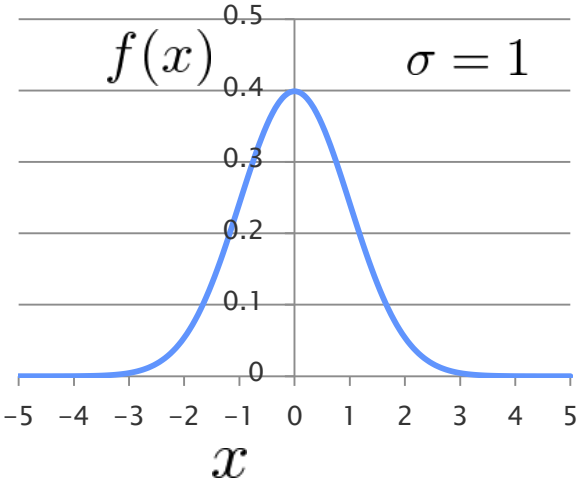
$f(x)\mathrm{d}x$ ：粒径が x と $x + \mathrm{d}x$ の間にある確率

規格化条件：$\int_{-\infty}^{\infty} f(x)dx = 1$

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。

演習課題

- 関数RAND()を用いて平均0の一様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様分布も）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の一様乱数を生成する。
  - 1列数千個のデータを12列作る。
- 12個の一様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、*Officeボタン/Excelのオプション/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/B49
…	…	…	…
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)
…	…

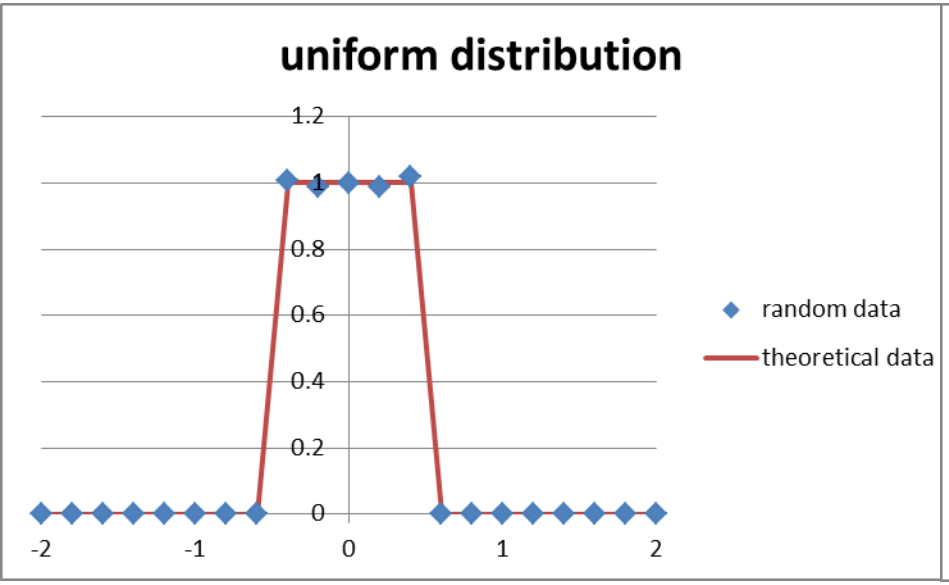
IF(条件, 値1, 値2)： 条件が真であれば値1を，偽であれば値2を返す。
AND(条件1, 条件2)： 条件1が真かつ条件2が真であれば真の値を返す。
したがって上の例は、セルC3の値が−0.5より大きく、0.5より小さいならば、セルに1が入力され、それ以外では0が入力される。

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

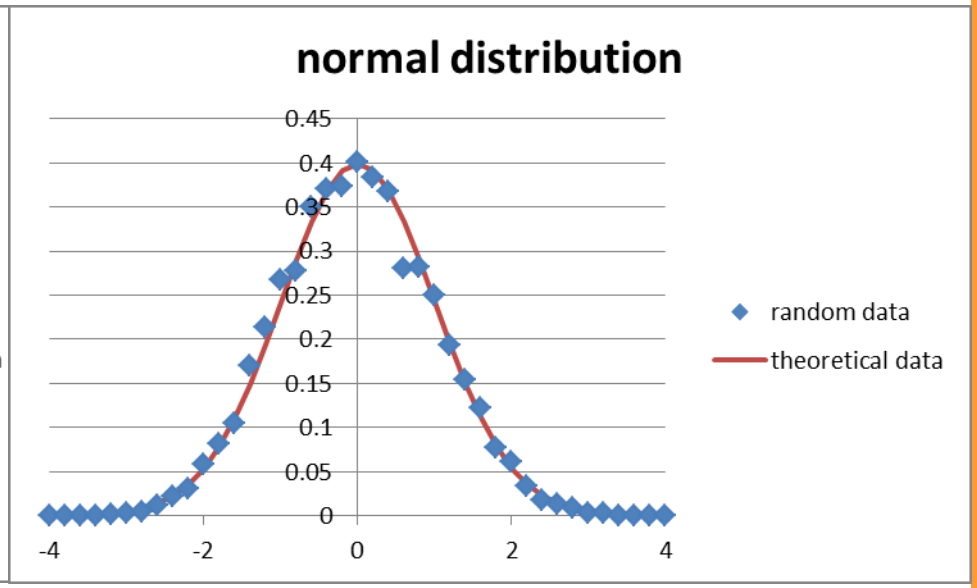
関数NORM.DIST($x, \bar{x}, \sigma^2$, FALSE) はxの値に対する平均$\bar{x}$ 分散σ^2 の正規分布 $f(x)$ の値を返す。

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。



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



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

参考

- （度数分布にする前の）データから平均値と分散を計算するには、関数 AVERAGE(範囲) と VAR.P(範囲) [Excel 2007 以前では VARP(範囲)] が使える（推定値のときは VAR.S または VAR）。
- 多数のデータに対し、関数を使わずに分散の計算をするときには、計算誤差に注意する必要がある。以下の式に従って計算すると計算誤差を小さくすることができる。

$$\sigma^2 = \frac{1}{N}\left\{\sum_{i=1}^{N}(x_i-\overline{x})^2 - \frac{1}{N}\left[\sum_{i=1}^{N}(x_i-\overline{x})\right]^2\right\}$$

推定値を計算する場合にはここの N を $N-1$ にする。