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

# 演習課題

- 最小2乗法によるデータのフィッティング
  - 最小2乗法の原理を理解する。
  - (実験値のかわりに)乱数を用いてデータを作成する。
  - 作成したデータをフィットする近似直線を求める。
- 頻度分布の作成
  - 頻度分布、分布関数について理解する。
  - 一様乱数を用いて、一様分布に従うデータおよび正規分布に従うデータを作成する。
  - 作成したデータの頻度分布を作成し、グラフにする。

# データのフィッティング（最小2乗法）

- 実験データ$(x_i, y_i)$を直線$y = ax + b$ でフィットする。直線からのずれの2乗和が最小になるように$a$, $b$ を決める。

$$f = \sum_i [y_i - y(x_i)]^2 = \sum_i [y_i - (ax_i + b)]^2$$

$$\frac{\partial f}{\partial a} = 0, \quad \frac{\partial f}{\partial b} = 0$$

⇩

$$\begin{pmatrix} \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i & \sum_i 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_i x_i y_i \\ \sum_i y_i \end{pmatrix}$$

これを解いて$a$, $b$ の値を求める。

# データの作成

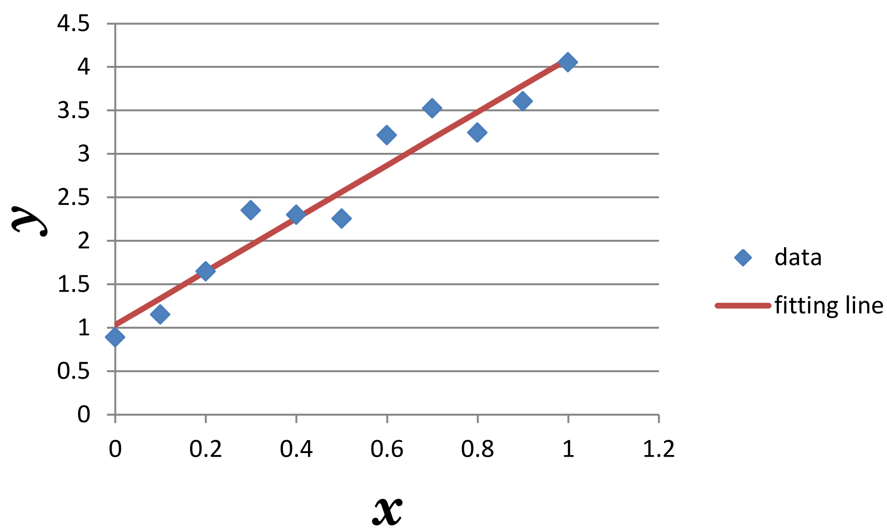
- 直線でフィットできるようなデータを人工的に作る。
  - 関数RAND()を使って乱数を発生させ、データにばらつきを与える。
  - RAND()は呼び出されるたびに、0と1の間のでたらめな数を返す関数である。
  - 例: $y = 3x + 1$ という直線関係が期待されるデータの作成

| x | y |
|---|---|
| 0 | = 3 * $A2 + 1 + RAND() − 0.5 |
| 0.1 | |
| 0.2 | |
| 0.3 | |
| | |

# フィッティング直線を求める

- $a$, $b$ に関する連立1次方程式の係数を求める。

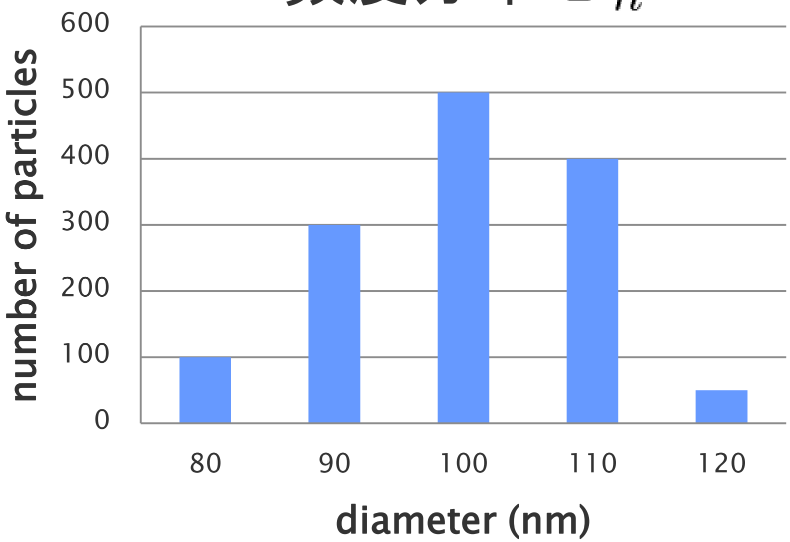
| x | y | x^2 | x*y |
|---|---|---|---|
| 0 | 0.707763 | 0 | 0 |
| 0.1 | 1.351046 | 0.01 | 0.135105 |
| 0.2 | 1.660645 | 0.04 | 0.332129 |
| 0.3 | 2.182343 | 0.09 | 0.654703 |
| 0.4 | 1.8433 | 0.16 | 0.73732 |
| 0.5 | 2.327833 | 0.25 | 1.163916 |
| 0.6 | 2.48472 | 0.36 | 1.490832 |
| 0.7 | 3.394455 | 0.49 | 2.376119 |
| 0.8 | 3.044672 | 0.64 | 2.435737 |
| 0.9 | 3.644913 | 0.81 | 3.280422 |
| 1 | 4.106025 | 1 | 4.106025 |
| | | | |
| | | | |
| X11= | 3.85 | Y1= | 16.71231 |
| X12= | 5.5 | Y2= | 26.74772 |
| X22= | 11 | Det= | 12.1 |
| | | | |
| a = (X22*Y1-X12*Y2)/Det | | | 3.034955 |
| b = (X11*Y2-X12*Y1)/Det | | | 0.914133 |



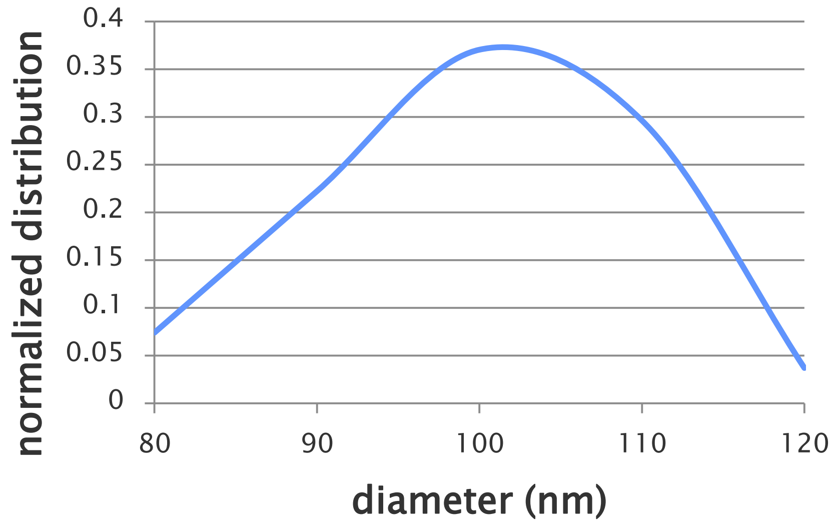
# 頻度分布

- 例 : 多数のコロイド粒子の粒径 $x_i \ (i = 1, 2, \ldots, N)$ の測定値から、頻度分布を作成する。

頻度分布 $F_n$




分布関数(確率密度) $f(x)$




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

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

$N \to \infty$

$\Delta x \to 0$

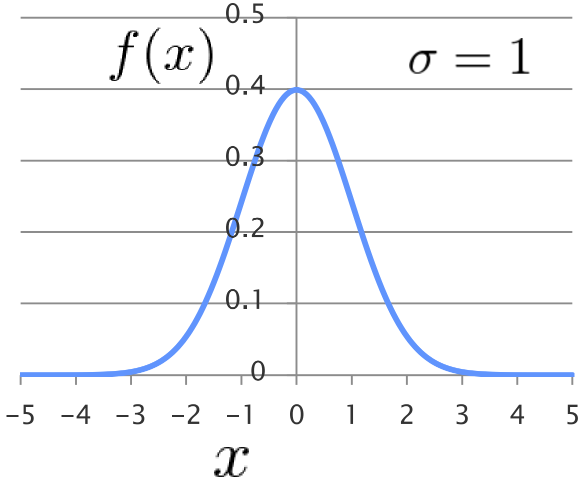
$f(x)\mathrm{d}x$ : 粒径が $x$ と $x + \mathrm{d}x$ の間にある確率

# 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？

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

($\sigma^2$ : 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

# データの作成

- RAND()を使って乱数を生成する（1万個くらい）。
  - RAND()は[0,1)の一様乱数を発生する。
- 一様乱数を12個足し合わせたものから6を引いて1つの乱数とし、それを1万個くらい生成する。

$$x = u_1 + u_2 + \cdots + u_{12} - 6$$

  - これは正規乱数（正規分布に従う乱数）を生成する簡便な方法として用いられる。

# 頻度分布の作成

- 分布関数 $f(x)$ に対して、$x$ の範囲と刻み幅 $\Delta x$ を決める。
- $x_n$ と $x_{n+1}$ の間に何個のデータがあるか COUNTIF を用いて数える。

セルE4の値以上の値をもつならば（カウントする）

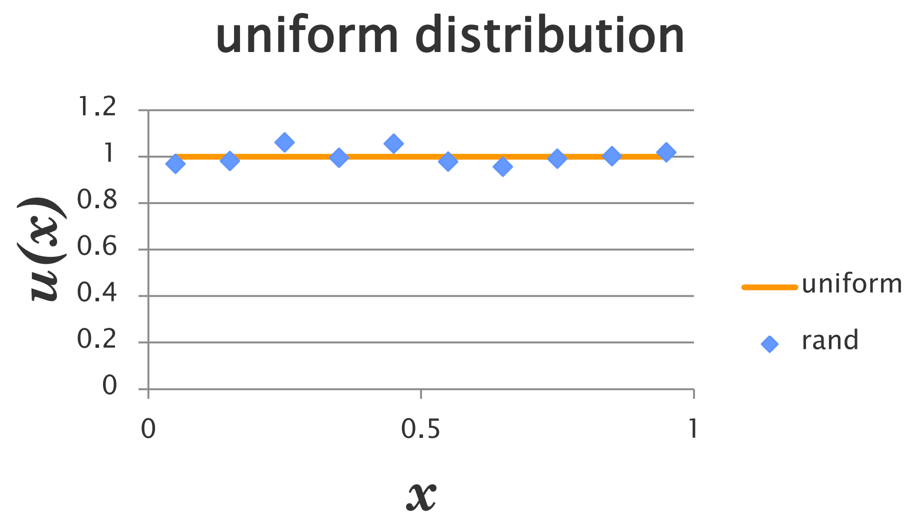
データの範囲　　　　データの範囲

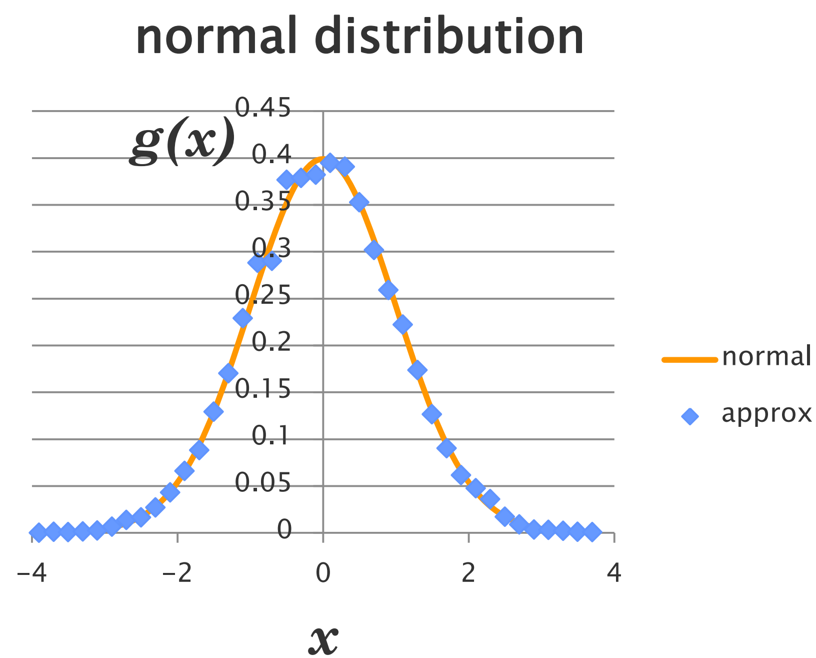
```
=COUNTIF($B$4:$B$10003, ">="&$E4) – COUNTIF($B$4:$B$10003, ">"&$E5)
```

| No. | uniform | Gauss | | X_n | X_c | frequency | normalized distribution | uniform |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.6387 | -0.63067 | | 0 | 0.05 | 1001 | 1.001 | 1 |
| 2 | 0.115474 | -0.45847 | | 0.1 | 0.15 | 1099 | 1.099 | 1 |
| 3 | 0.112641 | 0.841784 | | 0.2 | 0.25 | 987 | 0.987 | 1 |

# グラフの作成

- 一様分布のグラフ
  - 規格化された頻度
  - 分布関数
- 正規分布のグラフ
  - 規格化された頻度
  - 分布関数



# レポート

- 演習課題をレポートとしてA4用紙1枚にまとめ、学籍番号、氏名（自筆）を明記してこの時間内に提出。

- 最小2乗法によるデータのフィッティング
  - 作成したデータとフィッティング直線のグラフ
  - 期待される直線の式と最小2乗法によってえられた近似直線の式
- 頻度分布の作成
  - 一様分布の頻度データと分布関数のグラフ
  - 正規分布の頻度データと分布関数のグラフ