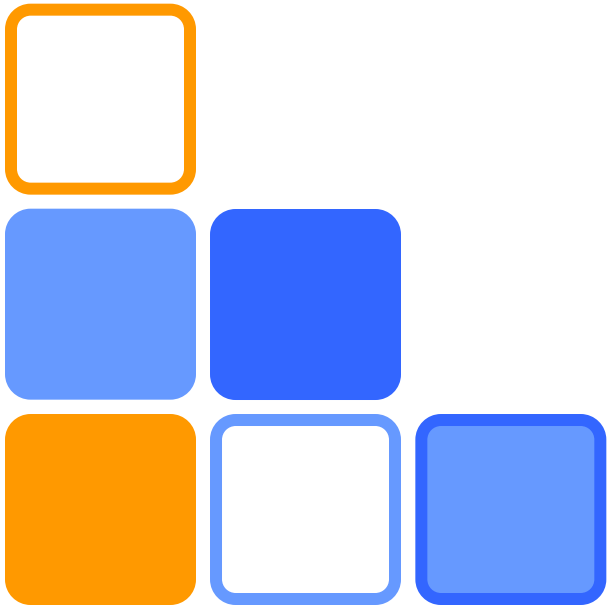
薬学情報処理演習　第2回

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

- 平均値と分散
  - N 個の(数値)データ $x_1, x_2, \cdots, x_N$
  - 平均値: $\bar{x} = \dfrac{x_1+x_2+\cdots+x_N}{N}$
  - 分散: $\sigma^2 = \dfrac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_N-\bar{x})^2}{N}$ (σ:標準偏差)
- 度数分布
  - 値$X_1, X_2, \cdots, X_n$ をとる度数(頻度): $F_1, F_2, \cdots, F_n$
  - $\bar{x} = \dfrac{F_1X_1+F_2X_2+\cdots+F_nX_n}{F_1+F_2+\cdots+F_n}$ $\qquad N = F_1 + F_2 + \cdots + F_n$
  - $\sigma^2 = \dfrac{F_1(X_1-\bar{x})^2+F_2(X_2-\bar{x})^2+\cdots+F_n(X_N-\bar{x})^2}{F_1+F_2+\cdots+F_n}$
- 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。

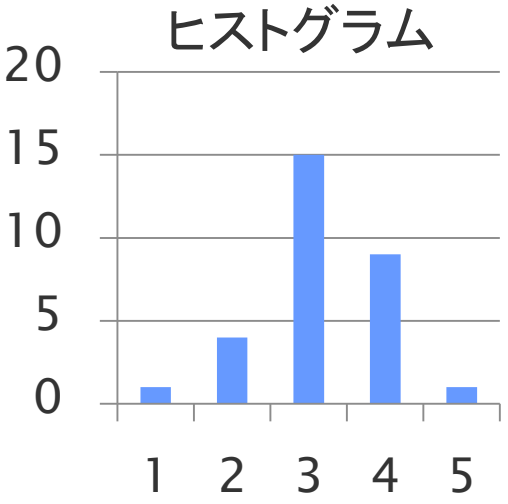
代表値 X_i	度数 F_i
1	1
2	4
3	15
4	9
5	1

データの分布

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

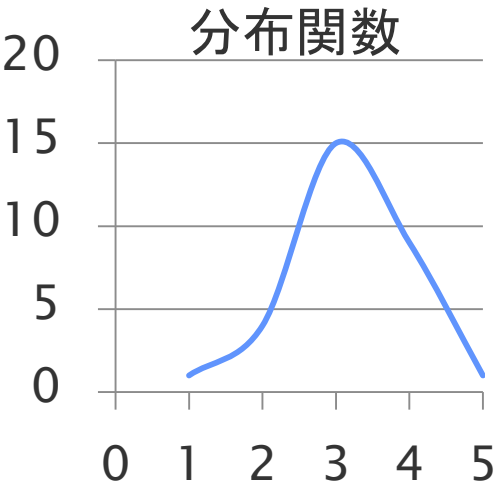
$$\bar{x} = \sum_{i=1}^{n} X_i f_i \qquad \sigma^2 = \sum_{i=1}^{n} (X_i - \bar{x})^2 f_i$$

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



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

2次元データの分布・平均値

- 2つの量(x_i, y_j)の度数分布f_{ij}

平均値：$\bar{x} = \sum_i \sum_j x_i f_{ij} / N$, $\bar{y} = \sum_i \sum_j y_j f_{ij} / N$

$(N = \sum_i \sum_j f_{ij})$

	140	150	160	170	180	190	$\Sigma_i f_{ij}$
40	1						1
50		1	3	1			5
60		1	1	2			4
70			1	2	3	1	7
80					2		2
90						1	1
$\Sigma_j f_{ij}$	1	2	5	5	5	2	

$$\bar{x} = \frac{140 \times 1 + 150 \times 2 + \cdots + 190 \times 2}{20} = 168.5 \simeq 170$$

$$\bar{y} = \frac{40 \times 1 + 50 \times 5 + \cdots + 90 \times 1}{20} = 63.5 \simeq 60$$

身長 (cm)	体重 (kg)
156	45
143	42
175	71
181	75
179	65
168	63
158	48
176	73
173	69
155	49
163	56
169	52
171	71
188	73
191	88
177	82
165	63
154	48
149	56
164	66
167.75	62.75

2次元データの相関

- 分散(variance)

$$\sigma_x^2 = \frac{1}{N}\sum_i \sum_j (x_i - \bar{x})^2 f_{ij} = \frac{1}{N}\sum_i \sum_j x_i^2 f_{ij} - \bar{x}^2$$

$$\sigma_y^2 = \frac{1}{N}\sum_i \sum_j (y_j - \bar{y})^2 f_{ij} = \frac{1}{N}\sum_i \sum_j y_j^2 f_{ij} - \bar{y}^2$$

- 共分散(covariance)

$$C_{xy} = \frac{1}{N}\sum_i \sum_j (x_i - \bar{x})(y_j - \bar{y}) f_{ij} = \frac{1}{N}\sum_i \sum_j x_i y_j f_{ij} - \bar{x}\bar{y}$$

- 相関係数(correlation coefficient)

$$r = \frac{C_{xy}}{\sigma_x \sigma_y} \qquad (|r| \leq 1)$$

r は x と y がどれくらい関連しているかを表す:

$r = 0$: 無相関
$r = 1$: 正の完全相関
$r = -1$: 負の完全相関

回帰直線

最小2乗法による直線近似

2つの量(x, y)の間に直線的な関係

$$y = ax + b \quad (a, b\text{は定数})$$

があると仮定し、直線からのずれの2乗和

$$Q = \sum_i \sum_j \left[y_j - (ax_i + b)\right]^2 f_{ij}$$

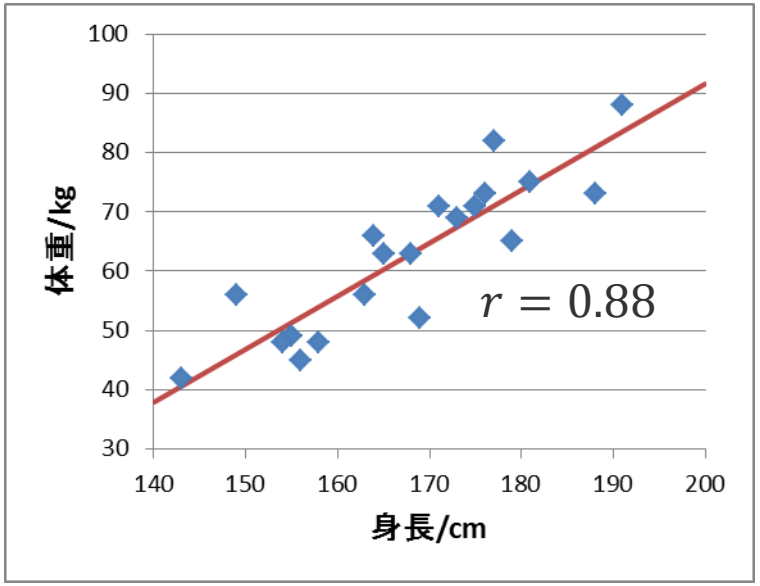
を最小にする a, b を求める:

$$\frac{\partial Q}{\partial a} = 0, \qquad \frac{\partial Q}{\partial b} = 0$$

回帰直線(regression line)

$$\frac{y - \bar{y}}{\sigma_y} = r\frac{x - \bar{x}}{\sigma_x}$$

が得られる。



データ $\{x_i, y_j\}$

↓

平均値 $\bar{x}, \bar{y}$ を計算

↓

分散σ_x^2, σ_y^2, 共分散 C_{xy}を計算

↓

相関係数 r を計算

↓

回帰直線 $y = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$ を得る

演習課題

- 関数RAND()を用いて相関をもつ擬似的な2次元データを作成する。
- 上記のデータから、平均値、分散、共分散、相関係数を計算し、回帰直線をデータと共にプロットする。
  - 回帰直線が予想したものに近いか確かめる。
  - データ長(N)を変えてみる。

データの作成

- 人工的に相関のあるデータを作る。疑似乱数を用いてばらつきのあるデータにする。
- 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、

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

である。

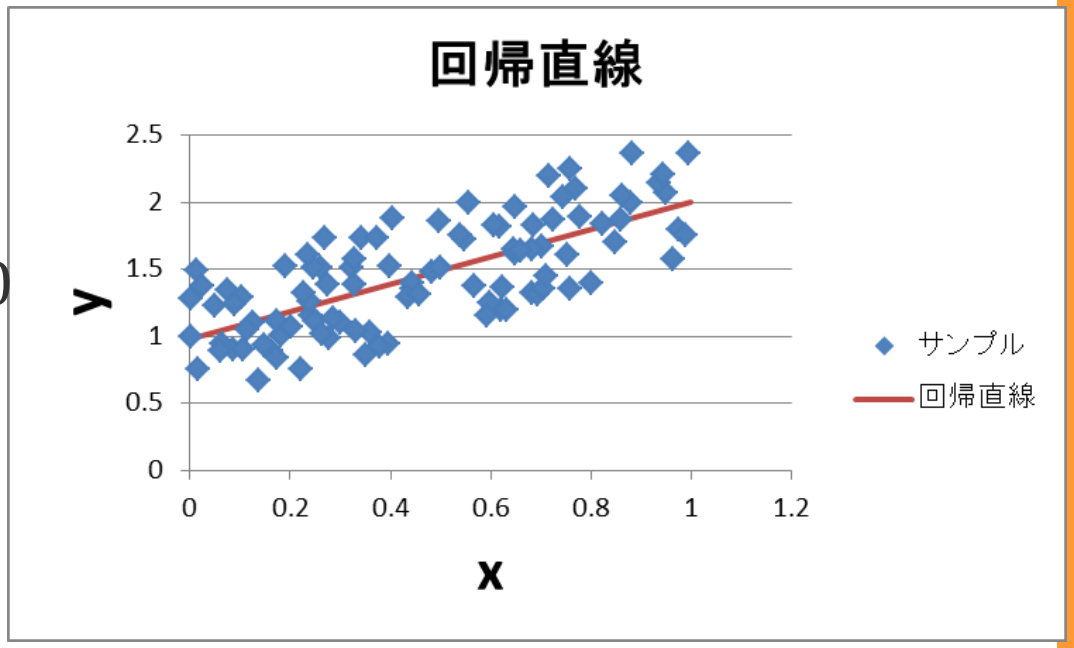
入力例

回帰分析						
パラメータ	a =	1				
	b =	1				
No.	x	y		sx2	sy2	Cxy
1	=rand()	=$C$2*B6+$C$3+RAND()-0.5			0.312885	0.167965
2	0.321401	1.512634		0.019534	0.003565	-0.00835
3	0.482437	1.479137		0.000453	0.000687	0.000558

99	0.58472	2.033889		0.001991	0.207663	0.020334
100	0.691387	2.095171		0.022888	0.267272	0.078213
平均値	0.540099	1.578188	分散	0.084319	0.175576	0.088404
			標準偏差	0.290378	0.419018	
			相関係数	0.726568		
			a=	1.048447	b=	1.011922

計算例

- 計算条件
  - $a = 1, b = 1$（予想される値）
  - データ長 $N = 100$
- 計算結果
  - $\bar{x} = 0.461, \bar{y} = 1.45$
  - $\sigma_x^2 = 0.0806, \sigma_y^2 = 0.160$
  - $C_{xy} = 0.0825, r = 0.726$
  - $a = 1.02, b = 0.980$



度数分布表の作成（参考）

- 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