2026 年度　データサイエンス数学基礎/統計学　（担当：松澤　寛）　レジメ

No.3　データの整理

$\sum$ 記号の定義とその性質

n 個の添え字がついた文字 $a_1, a_2, \cdots, a_n$ に対して

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \cdots + a_n$$

と表す.
$\sum$ 記号の性質：

- 定数 c に対して $\sum_{i=1}^{n} c = c + c + \cdots + c = nc$
- $$\sum_{i=1}^{n}(a_i + b_i) = (a_1 + b_1) + (a_2 + b_2) + \cdots + (a_n + b_n)$$
$$= (a_1 + a_2 + \cdots + a_n) + (b_1 + b_2 + \cdots + b_n) = \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i$$
- 定数 c に対して
$$\sum_{i=1}^{n} ca_i = ca_1 + ca_2 + \cdots + ca_n = c(a_1 + a_2 + \cdots + a_n) = c\sum_{i=1}^{n} a_i$$

度数分布表とヒストグラム

- 学生 100 人の試験結果

100, 49, 63, 38, 72, 33, 64, 67, 52, 28, 85, 72, 64, 61, 91, 62, 84
61, 47, 68, 59, 82, 49, 69, 46, 84, 72, 17, 70, 56, 72, 54, 46, 81
100, 71, 38, 65, 55,100, 43, 49, 91, 86, 76, 47,100, 56, 51, 53, 50
44, 83, 63, 55, 46, 30, 11, 57, 72, 53, 71, 72, 59, 38, 50, 18, 40
100, 87, 71, 43, 18, 75, 90, 36, 42, 91, 52, 61, 42, 50, 49, 81, 59
67, 54, 58, 69, 77, 82, 15, 29, 66, 65, 68, 55, 33, 71, 45,

- 一般に 10 点刻みのような等分割を考え，1 つ 1 つの区間を**階級**という．各階級に存在するデータの数を**度数**という．全ての階級に対して階級と度数の対応をまとめた次のような表を**度数分布表**という.

点数	0-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
人数	0	5	3	7	18	17	18	13	11	8

- 度数を次のようなグラフとしてまとめたものを**ヒストグラム**という.

データを代表する量

- **平均**：n 個のデータ $x_1, x_2, \cdots. x_n$ に対して

$$\frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

を**平均**といい，$\overline{x}$ で表す.

- 各データの 2 乗の値 $x_1^2, x_2^2, \cdots, x_n^2$ の平均

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2$$

を $\overline{x^2}$ と表す.

- **平均の性質**： n 個のデータ $x_1, x_2, \cdots, x_n$ $y_1, y_2, \cdots, y_n$ の間 $y_i = ax_i + b$ $(a, b:$ 定数$)$ の関係があるとき

$$\overline{y} \quad = \quad \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) = a\frac{1}{n}\sum_{i=1}^{n} x_i + b = a\overline{x} + b$$

が成り立つ.

- **例**：5 人の学生の身長 $x_1, x_2, \cdots, x_5$ が

$$165.0,\ 170.2,\ 168.3,\ 175.3,\ 162.3$$

であったとする．$y_i = x_i - 165.0$ とすると $y_1, y_2, \cdots, y_5$ は

$$0.0, 5.2, 3.3, 10.3, -2.7$$

である．したがって $3.22 = \overline{y} = \overline{x} - 165.0$ つまり $\overline{x} = 168.22$ である．

- **中央値**：データを大きさの順に並べたとき，中央に位置する値．データが奇数個，つまり $2n+1$ 個あるとき $x_1, \cdots, x_{2n+1}$ （$x_1 \leq x_2 \leq \cdots \leq x_{2n+1}$）の中央値は x_{n+1} である．偶数個，つまり $2n$ 個あるとき，$x_1, \cdots, x_n$ $(x_1 \leq x_2 \leq \cdots \leq x_{2n-1} \leq x_{2n})$ の中央値は $\dfrac{1}{2}(x_n + x_{n+1})$ である．

  例：8 人の学生の所持金が以下のように

  $$1000 \quad 1500 \quad 1500 \quad 2000 \quad 2000 \quad 3000 \quad 4000 \quad 20000$$

  であったとき，平均は 4375 であり，中央値は 2000 である．

- **最頻値**：度数分布表で最も度数が高い階級の中央値という．上のテストの例では 45 と 65 である．

  注意 ： 最頻値は階級の分け方に依存する．

データのばらつきを表す量

- **範囲**：最大値 − 最小値
- **分散**：$v_x = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- **分散の計算**

$$\begin{aligned} v_x &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}((x_i)^2 - 2\overline{x}x_i + (\overline{x})^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - 2\overline{x}\frac{1}{n}\sum_{i=1}^{n}x_i + (\overline{x})^2 = \overline{x^2} - 2(\overline{x})^2 + (\overline{x})^2 \\ &= \overline{x^2} - (\overline{x})^2 \end{aligned} \tag{1}$$

- **分散の性質**：データ $x_1, \cdots, x_n,\ y_1, \cdots, y_n$ が，$y_i = ax_i + b$ $(a, b:$ 定数$)$ の関係があるとき，$\overline{y} = a\overline{x} + b$ より

$$\begin{aligned} v_y &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}\{(ax_i + b) - (a\overline{x} + b)\}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(ax_i - a\overline{x})^2 = a^2\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= a^2 v_x \end{aligned} \tag{2}$$

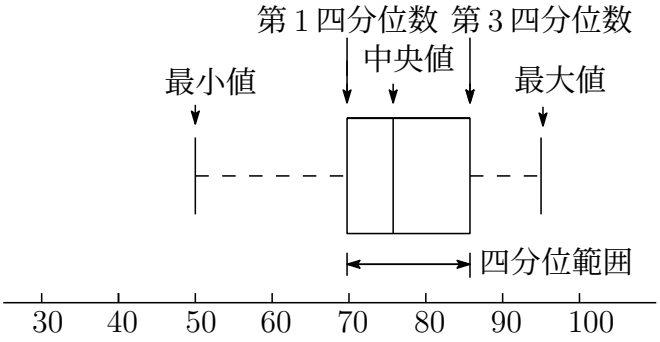
- **標準偏差**：$\sigma_x = \sqrt{v_x}$

箱ひげ図 何組かの同種のデータ（例えば，同じ学年の模試のデータ）を比較するときに用いる．

- 最大値と最小値のみに依存するので極端に他のデータから離れたデータがあるとき，それ以外のデータの散らばり具合を表していない．中央の 50%のデータの範囲を図る**四分位法**
- データを昇順（小 $\to$ 大）に並べ，データを 4 分割
- 左から第 1，2，3 分割点をそれぞれ**第 1 四分位数**，**第 2 四分位数**，**第 3 四分位数**という．**第 2 四分位数は中央値**である．
- **第 3 四分位数 − 第 1 四分位数** : **四分位範囲**
  - 第 1 四分位数：最小値から中央値の間にあるデータの中での中央値
  - 第 3 四分位数：中央値から最大値の間にあるデータの中での中央値
- **第 1 四分位数 − 四分位範囲 $\times$1.5** より小さいデータと
  第 3 四分位数 + 四分位範囲 $\times$1.5 より大きいデータを**外れ値**という．
- データの広がりは**箱ひげ図**で表す．

例 クラス A: 72 78 68 71 80 90 88 95 85 75 60 50

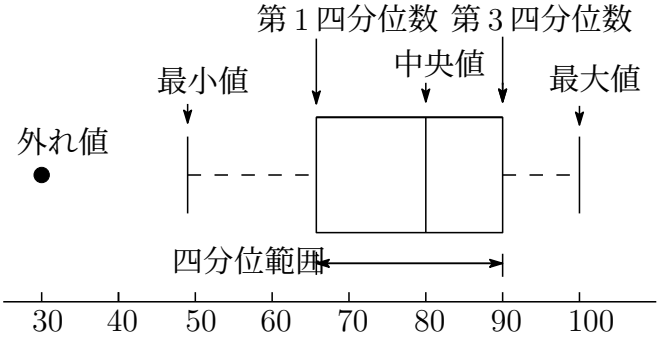
- 昇順に並べ, 4 分割 : 50 60 68 | 71 72 75 | 78 80 85 | 88 90 95
- 中央値（第 2 四分位数）: $\dfrac{75+78}{2} = 76.5$
- 第 1 四分位数 : $\dfrac{68+71}{2} = 69.5$，第 3 四分位数 : $\dfrac{85+88}{2} = 86.5$
- 四分位範囲 : $86.5 - 69.5 = 17$
- 第 1 四分位数 − 四分位範囲 $\times 1.5 = 69.5 - 17 \times 1.5 = 44.0$,
- 第 3 四分位数 + 四分位範囲 $\times 1.5 = 86.5 + 17 \times 1.5 = 112.0$

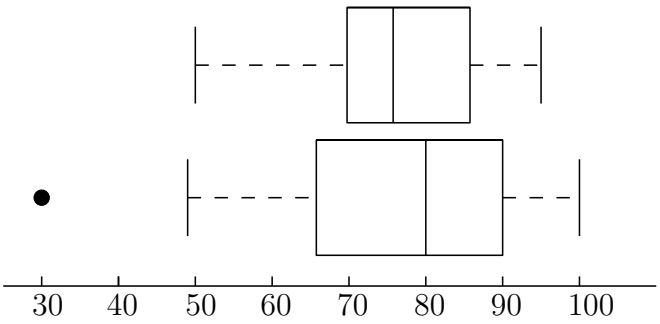


例

クラス B: 70 84 88 65 76 92 100 30 98 85 68 48

- 昇順に並べ, 4 分割：30 48 65 | 68 70 76 | 84 85 88 | 92 98 100
- 中央値（第 2 四分位数）：$\dfrac{76+84}{2} = 80.0$
- 第 1 四分位数：$\dfrac{65+68}{2} = 66.5$，第 3 四分位数：$\dfrac{88+92}{2} = 90.0$
- 四分位範囲：$90.0 - 66.5 = 23.5$
- 第 1 四分位数 $-$ 四分位範囲 $\times 1.5 = 66.5 - 23.5 \times 1.5 = 31.25$,
- 第 3 四分位数 $+$ 四分位範囲 $\times 1.5 = 90.0 + 23.5 \times 1.5 = 125.25$
- **30 が外れ値**
- 外れ値がある場合，外れ値を除いたデータの最大値，最小値を示し，外れ値は個別に示す（外れ値を除いたデータの最小値は 48）





- 同じ種類の複数のデータの箱ひげ図を重ねることができる．箱ひげ図は，複数グループのデータの分布を比較するときにヒストグラムより便利であることが多い．