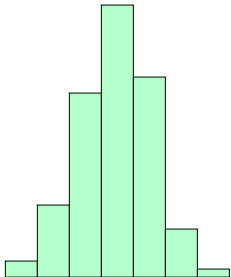


§ 1.3 散布度

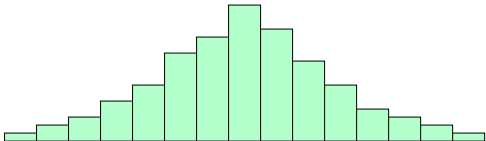
データの要素が実数で表される量であるとする．このとき、データの要素がばらついているのか集中しているのかということがしばしば問題になる．

データの要素のばらつきの程度を示す数量をデータの散布度という．つまり、データの散布度が大きいということはデータの要素がばらついている（散らばっている）ことであり、データの散布度が小さいということはデータの要素が集中していることである．

散布度の大小をヒストグラムで表すと例えば次のようになる．2つのヒストグラムにおいて階級幅は同じとする．



散布度が比較的小さい状態の例



散布度が比較的大きい状態の例

正の自然数 n に対して n 個の数 $x_1, x_2, x_3, \dots, x_n$ を要素とするデータについて、平均を $\bar{x}$ とおく：

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n} \sum_{k=1}^n x_k .$$

このデータの各要素 x_k ($k = 1, 2, 3, \dots, n$) に対して、平均 $\bar{x}$ との差 $x_k - \bar{x}$ を x_k の偏差という．各々の要素がこの平均 $\bar{x}$ より離れれば離れるほど散布度は大きくなるはずである．つまり、各要素 x_k ($k = 1, 2, 3, \dots, n$) と平均 $\bar{x}$ の距離 $|x_k - \bar{x}|$ が大きければ大きいほど散布度は大きくなるはずである．偏差の絶対値 $|x_k - \bar{x}|$ は数学的に扱いにくいので、代わりに偏差の2乗 $(x_k - \bar{x})^2$ をよく用いる．

散布度として**分散** (variance) と**標準偏差** (standard deviation⁴⁾) とがよく使われる．

[定義] 正の自然数 n に対して n 個の数 $x_1, x_2, x_3, \dots, x_n$ を要素とするデータについて、平均を $\bar{x}$ とおく：

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n} \sum_{k=1}^n x_k .$$

データの各要素の偏差の平方の総和

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2 = \sum_{k=1}^n (x_k - \bar{x})^2$$

を偏差平方和という．偏差平方和 $\sum_{k=1}^n (x_k - \bar{x})^2$ を要素の個数 n で割った

$$\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 .$$

を分散という．分散 $\frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2$ の0以上の平方根 $\sqrt{\frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2}$ を標準偏差という．

データの分散について次の定理が成り立つ．

[定理1.3] 正の自然数 n に対して n 個の数 $x_1, x_2, x_3, \dots, x_n$ を要素とするデータについて、平均を $\bar{x}$ とおくと、分散 v は

$$v = \frac{1}{n} \sum_{k=1}^n x_k^2 - \bar{x}^2 .$$

[証明] データの平均 $\bar{x}$ は $\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$ なので $\sum_{k=1}^n x_k = n\bar{x}$ ．データの分散 v は

$$\begin{aligned} v &= \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 = \frac{1}{n} \sum_{k=1}^n (x_k^2 - 2x_k\bar{x} + \bar{x}^2) \\ &= \frac{1}{n} \left\{ \sum_{k=1}^n x_k^2 - 2\left(\sum_{k=1}^n x_k\right)\bar{x} + \sum_{k=1}^n \bar{x}^2 \right\} \\ &= \frac{1}{n} \left(\sum_{k=1}^n x_k^2 - 2n\bar{x}\bar{x} + n\bar{x}^2 \right) = \frac{1}{n} \left(\sum_{k=1}^n x_k^2 - n\bar{x}^2 \right) \\ &= \frac{1}{n} \sum_{k=1}^n x_k^2 - \bar{x}^2 . \end{aligned}$$

証明終了

正の自然数 n に対して n 個の数 $x_1, x_2, x_3, \dots, x_n$ を要素とするデータについて、平均を $\bar{x}$ とおくと、分散 v は、定義では

$$v = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2$$

であるが、この式で分散 v を計算するよりも、上述の定理の式

$$v = \frac{1}{n} \sum_{k=1}^n x_k^2 - \left(\frac{1}{n} \sum_{k=1}^n x_k \right)^2$$

で計算の方が簡単なことがある．

不偏分散 (unbiased variance) と言われる散布度もある．

[定義] 2以上の自然数 n に対して n 個の数 $x_1, x_2, x_3, \dots, x_n$ を要素とするデータについて、平均を $\bar{x}$ とおく：

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n} \sum_{k=1}^n x_k .$$

偏差平方和 $\sum_{k=1}^n (x_k - \bar{x})^2$ を $n-1$ で割った

$$\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2$$

を不偏分散という．

不偏分散において偏差平方和 $\sum_{k=1}^n (x_k - \bar{x})^2$ を $n-1$ で割る理由は第6章第3節で述べる．

[例題] ある4名の学生の身長を測って次のデータが得られたとする：

162cm, 180cm, 156cm, 170cm ．

この4名の学生の身長の分散と不偏分散と標準偏差とを求める．

4名の学生の身長の平均は

$$\frac{1}{4}(162 + 180 + 156 + 170)\text{cm} = 167\text{cm} .$$

4名の学生の身長の分散は

$$\begin{aligned} \frac{1}{4}\{(162 - 167)^2 + (180 - 167)^2 + (156 - 167)^2 + (170 - 167)^2\}\text{cm}^2 &= \frac{324}{4}\text{cm}^2 \\ &= 81\text{cm}^2 . \end{aligned}$$

4名の学生の身長の不偏分散は

$$\begin{aligned} \frac{1}{3}\{(162 - 167)^2 + (180 - 167)^2 + (156 - 167)^2 + (170 - 167)^2\}\text{cm}^2 &= \frac{324}{3}\text{cm}^2 \\ &= 108\text{cm}^2 . \end{aligned}$$

4名の学生の身長の標準偏差は

$$\sqrt{81\text{cm}^2} = 9\text{cm} .$$

終

[問1.3] ある4名の学生の体重が以下のものであったとする．

60kg, 56kg, 71kg, 65kg ．

これら4名の学生の体重の分散と不偏分散と標準偏差とを求めよ．

データの散布度としては、分散・標準偏差以外にも範囲（レンジ） (range) といわれるものがある．データの範囲（レンジ）とは、データの中の最大値から最小値を引いた結果である．

⁴⁾ SD と略記されることがある．