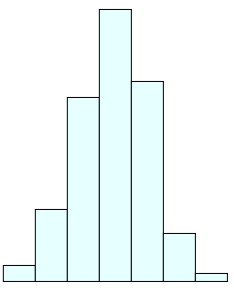
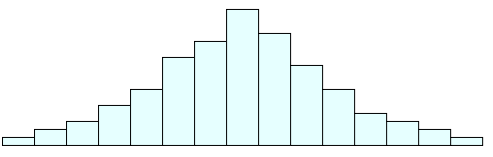
§1.3　散布度

データの要素が実数で表される量であるとします．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もしばしば問題になります．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います．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す．

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ります．以下の 2 つのヒストグラムにおいて階級幅は同じとします．



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 $\overline{x}$ とおきます：

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

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

散布度として分散と標準偏差とがよく使われます．

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 $\overline{x}$ とおく：

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

データの各要素の偏差の平方の総和

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

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

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

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

不偏分散と言われる散布度もあります．

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 $\overline{x}$ とおく：

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

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

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

を不偏分散という．

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

データの分散について次の定理が成り立ちます．

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の分散 v は

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

証明　データの平均を $\overline{x}$ とおく： $\overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$ ．分散 v の定義は $\frac{1}{n}\sum_{k=1}^{n} (x_k - \overline{x})^2$ なので，

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

（証明終り）

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 $\overline{x}$ とおくとき，分散 v は，定義では

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

ですが，この式で分散 v を計算するよりも，上述の定理の式

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

で計算する方が簡単なことがあります．

例題 1.3　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：

$$162\,\mathrm{cm},\quad 180\,\mathrm{cm},\quad 156\,\mathrm{cm},\quad 170\,\mathrm{cm} \ .$$

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

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

4 名の学生の身長の分散は

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

4 名の学生の身長の不偏分散は

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

4 名の学生の身長の標準偏差は

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

終

問題 1.3　ある 4 名の学生の体重が以下のようであったとします．

$$60\,\mathrm{kg},\quad 56\,\mathrm{kg},\quad 71\mathrm{kg},\quad 65\mathrm{kg} \ .$$

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なさい．

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ります．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ものです．