## 5章　データの分析

神の御心を知るには
統計学を学ばねばならない。

イギリスの看護師，近代看護教育の生みの親。
クリミア戦争に看護師として従軍し，兵士の死亡データを集計・分析した。
その結果をもとにそれまで劣悪だった戦地の病院の環境を改善したところ，負傷した兵士の死亡率が大きく下がった。
また，国ごとにバラバラであった死亡統計の取り方に対し，集計方法を規格化して
国際的に統一されたものにするよう提案し，医療の技術向上に貢献した。

# 1 節 データの整理と分析

## 1 データの整理

**データ**

集められた資料から集団の特徴や傾向をとらえるためには，資料をどのように整理し，表やグラフに表せばよいのかを考えてみよう。

次の資料は，あるクラスの「一か月の読書時間」について，電車・バス通学の A 班と徒歩・自転車通学の B 班に分けて，調べた結果である。

| A 班 20 人（単位　時間） | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 10 | 7 | 14 | 5 | 9 | 15 | 0 | 13 | 18 |
| 0 | 8 | 11 | 10 | 15 | 19 | 6 | 23 | 9 | 5 |

| B 班 15 人（単位　時間） | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 20 | 6 | 0 | 14 | 16 | 23 | 1 | 4 | 5 | 0 |
| 18 | 13 | 21 | 0 | 9 | | | | | |

このような資料を**データ**という。また，読書時間のように，データの特性を表す数量を**変量**という。

**データの整理**

右の表は，A 班の読書時間の結果をもとに，0 時間から 24 時間までの間を 4 時間ずつの区間に分け，その区間に入っている人数を調べてまとめたものである。

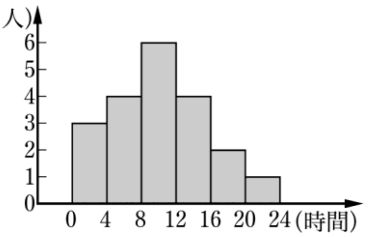
| 読書時間（時間） | 度数 |
|---|---|
| 以上～未満 | |
| 0 ～ 4 | 3 |
| 4 ～ 8 | 4 |
| 8 ～ 12 | 6 |
| 12 ～ 16 | 4 |
| 16 ～ 20 | 2 |
| 20 ～ 24 | 1 |
| 計 | 20 |

このように，データを整理するために用いる区間を**階級**，区間の幅を**階級の幅**，階級の真ん中の値を**階級値**という。

また，それぞれの階級に入っているデータの値の個数をその階級の**度数**，各階級に度数を対応させたものを**度数分布**，それを表にしたものを**度数分布表**という。

**問 1**　B 班の読書時間の度数分布表を作成せよ。

度数分布をグラフにした図が**ヒストグラム**である。A 班の読書時間のヒストグラムは右の図のようになる。




度数分布表やヒストグラムを用いるとデータの分布が見やすくなる。

**問 2** B 班の読書時間のヒストグラムをかけ。

**相対度数**

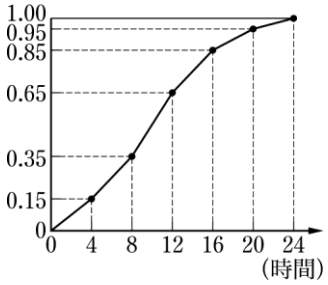
A 班と B 班は人数が異なるから，度数を見るだけではデータを比較しにくい。そこで，度数の代わりに，各階級の度数を度数の合計で割った値を用いるとよい。この値を**相対度数**という。

すなわち　　**相対度数** $=\dfrac{\text{その階級の度数}}{\text{度数の合計}}$

相対度数を用いることで，ある階級の度数が全体に占める割合がわかる。

また，相対度数を小さい階級からその階級の値まで合計して得られる値を**累積相対度数**という。下の表は，A 班の読書時間の度数分布表に相対度数と累積相対度数を並べて記入したものである。さらに，各階級の累積相対度数を折れ線でつないだものを**累積相対度数折れ線**という。

| 読書時間（時間） | 度数 | 相対度数 | 累積相対度数 |
|---|---|---|---|
| 以上～未満 | | | |
| 0 ～ 4 | 3 | 0.15 | 0.15 |
| 4 ～ 8 | 4 | 0.20 | 0.35 |
| 8 ～ 12 | 6 | 0.30 | 0.65 |
| 12 ～ 16 | 4 | 0.20 | 0.85 |
| 16 ～ 20 | 2 | 0.10 | 0.95 |
| 20 ～ 24 | 1 | 0.05 | 1.00 |
| 計 | 20 | 1.00 | |




A班の読書時間の
累積相対度数折れ線

**問 3** 問 1 で作成した度数分布表に，B 班の読書時間の相対度数と累積相対度数を付け加えよ。

## 2 代表値

いくつかのデータを比べるとき,それぞれのデータの特徴を1つの数値で表すと比較しやすい。そのような数値を**代表値**という。

代表値としては，平均値，中央値，最頻値がよく知られている。

**平均値**

$x$ を変量とし，データの $n$ 個の値 $x_1$，$x_2$，…，$x_n$ が与えられているとき

$$\frac{1}{n}(x_1+x_2+\cdots+x_n) \qquad \text{平均値}=\frac{\text{データの値の総和}}{\text{データの値の個数}}$$

をデータの**平均値**といい，記号 $\overline{x}$ で表す。

**例 1** 158ページのA班の読書時間の平均値を求めてみよう。

$$\frac{1}{20}(3+10+7+\cdots+23+9+5)=\frac{1}{20}\cdot 200=10\ (\text{時間})$$

**問 4** 158ページのB班の読書時間の平均値を求めよ。

次に，度数分布から平均値を求める方法を考えてみよう。右の表は，変量 $x$ の度数分布表である。このとき，平均値 $\overline{x}$ は各階級に含まれるデータの値がすべてその階級値に等しいとみなして，次の式で計算する。

| 階級値 | 度数 |
|---|---|
| $x_2$ | $f_2$ |
| $x_1$ | $f_1$ |
| ⋮ | ⋮ |
| $x_r$ | $f_r$ |
| 計 | $n$ |

$$\overline{x}=\frac{1}{n}(x_1f_1+x_2f_2+\cdots+x_rf_r)$$

**例 2** 158ページのA班の読書時間の平均値を度数分布表から求めてみよう。右の表より

| 階級値 | 度数 |
|---|---|
| 2 | 3 |
| 6 | 4 |
| 10 | 6 |
| 14 | 4 |
| 18 | 2 |
| 22 | 1 |
| 計 | 20 |

$$\frac{1}{20}(2\cdot 3+6\cdot 4+10\cdot 6+14\cdot 4+18\cdot 2+22\cdot 1)$$

$$=\frac{1}{20}\cdot 204=10.2\ (\text{時間})$$

**注意** 例1と例2のように，データから直接求めた平均値と度数分布表から求めた平均値では，値が異なることがある。

**問 5** 158 ページの問 1 で作成した度数分布表を用いて，B 班の読書時間の平均値を小数第 2 位を四捨五入して，小数第 1 位まで求めよ。

**中央値**

データのすべての値を小さい順に並べたとき，中央の順位にくる値を**中央値**または**メジアン**という。ただし，データの値の個数が偶数 $2n$ 個のときは，第 $n$ 番目と第 $n+1$ 番目のデータの値の平均値を中央値とする。

**例 3** 158 ページの A 班の読書時間の中央値を求めてみよう。

20 個のデータの値を小さい方から順に並べると

0　0　3　5　5　6　7　8　9　9　10　10　11　13　14　15　15　18　19　23

となる。このデータの中央値は，10 番目の値 9 と 11 番目の値 10 の平均値　$\frac{1}{2}(9+10)=9.5$ (時間)

である。

**問 6** 158 ページの B 班の読書時間の中央値を求めよ。

**最頻値**

データを度数分布表に整理したとき，度数が最も多い階級の階級値を**最頻値**または**モード**という。(*)最頻値はヒストグラムにおいて，最も多い度数を表す長方形における階級値である。

**例 4** 158 ページの A 班の読書時間の最頻値を求めてみよう。

度数分布表で，読書時間の度数が最も多い階級は 8 時間以上 12 時間未満である。

よって，最頻値はこの階級の階級値を求めて，10 時間である。

**問 7** 158 ページの問 1 で作成した度数分布表を用いて，B 班の読書時間の最頻値を求めよ。

---

(*)データの中で最も多く出てくる値を最頻値ということもある。

## 3 箱ひげ図

**四分位数**

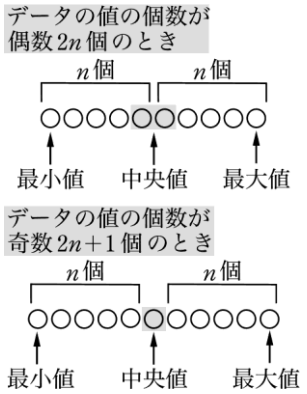
データの特徴をよりくわしく表すために，データの値を小さい方から順に並びかえて，次のような数値を考えてみよう。

① データの中央値を求める。

② 右の図のように，中央値を境にしてデータの値の個数が等しくなるように 2 つの部分に分ける。

③ 2 つに分けたうち，最小値を含む方のデータの中央値を求める。

④ 2 つに分けたうち，最大値を含む方のデータの中央値を求める。




小さい順に，③の値を**第 1 四分位数**，①の値を**第 2 四分位数**，④の値を**第 3 四分位数**といい，それぞれ $Q_1$, $Q_2$, $Q_3$ で表す。これらを合わせて**四分位数**という。3 つの四分位数 $Q_1$, $Q_2$, $Q_3$ は，データの値の小さい方から 25%, 50%, 75% に対応する数値であるともいえる。

**例 5** 158 ページの A 班の読書時間の四分位数を求めてみよう。

20 個のデータの値を小さい方から順に並べると，次のようになる。

$\underline{0\quad 0\quad 3\quad 5\quad 5\quad 6\quad 7\quad 8\quad 9\quad 9}\ \underline{10\quad 10\quad 11\quad 13\quad 14\quad 15\quad 15\quad 18\quad 19\quad 23}$

このデータの第 2 四分位数は，例 3 より　$Q_2 = 9.5$ (時間)

第 1 四分位数は　$\underline{0\quad 0\quad 3\quad 5\quad 5\quad 6\quad 7\quad 8\quad 9\quad 9}$

の中央値より　$Q_1 = \frac{1}{2}(5+6) = 5.5$ (時間)

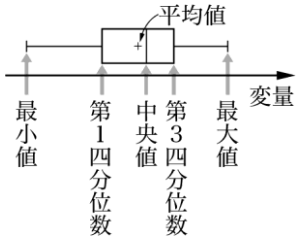
第 3 四分位数は　$\underline{10\quad 10\quad 11\quad 13\quad 14\quad 15\quad 15\quad 18\quad 19\quad 23}$

の中央値より　$Q_3 = \frac{1}{2}(14+15) = 14.5$ (時間)

**問 8** 158 ページの B 班の読書時間の四分位数を求めよ。

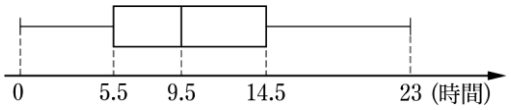
**箱ひげ図**

データの分布を最小値，第1四分位数，中央値，第3四分位数，最大値の5つの数値を用いて要約する方法がある。これを**5数要約**という。右の図は，5数要約を箱と線（ひげ）を用いて表したものであり，**箱ひげ図**という。(*)




**例 6** 158ページのA班の読書時間の箱ひげ図をかいてみよう。

例5より，箱ひげ図は次の図のようになる。

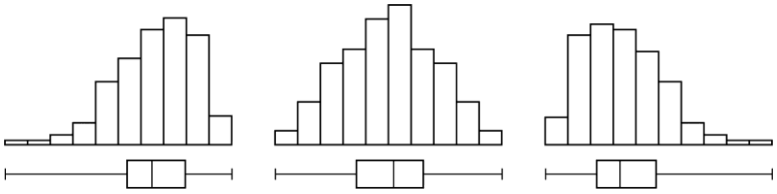



**問 9** 158ページのB班の読書時間の箱ひげ図をかけ。

**箱ひげ図とヒストグラム**

箱ひげ図はヒストグラムと同様に，データの分布を表現するのに適している。ヒストグラムでは，度数分布表のすべての階級の度数が必要であるのに対して，箱ひげ図は最小値，第1四分位数，中央値，第3四分位数，最大値の5つの数値がわかると，かくことができる。

箱ひげ図とヒストグラムの関係について考えてみよう。



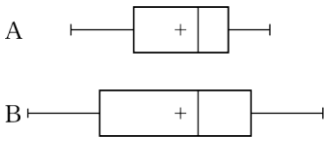
上の図のように，1つの山のヒストグラムの場合，ヒストグラムの山の高い部分に箱ひげ図の箱が対応し，山のすその部分に箱ひげ図のひげが対応している。

---

(*)箱ひげ図に平均値の位置を表すこともある。

## 4 箱ひげ図とデータの散らばり

右の図は，データの値の個数も平均値も中央値も等しい 2 つのデータ A, B の箱ひげ図である。




このように，平均値や中央値は等しくても，データの分布のようすが大きく異なることがある。箱ひげ図を用いることによって，データの分布のようすを視覚的に比較することを考えてみよう。

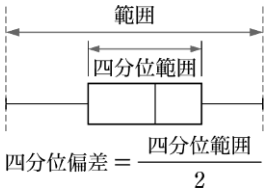
### 範囲と四分位偏差

あるデータにおいて，データの最大値から最小値を引いた値をそのデータの分布の**範囲**または**レンジ**という。

範囲は，データのすべての値を含む大きさを表しているから，極端にはずれた値の影響を受ける。

そこで，データの散らばり具合いを四分位数をもとにして，より的確に表すことを考えてみよう。

第 3 四分位数から第 1 四分位数を引いた値を**四分位範囲**といい，四分位範囲を 2 で割った値を**四分位偏差**という。




$$\text{四分位偏差} = \frac{\text{四分位範囲}}{2}$$

**例 7** 158 ページの A 班の読書時間の範囲，四分位範囲，四分位偏差を求めてみよう。

範囲は $23 - 0 = 23$ (時間)

四分位範囲は $14.5 - 5.5 = 9$ (時間)

四分位偏差は $\frac{9}{2} = 4.5$ (時間)

となる。

**問 10** 158 ページの B 班の読書時間の範囲，四分位範囲，四分位偏差をそれぞれ求めよ。

下の図1は，ある年の東京における各月の日ごとの平均気温の平均値（単位 ℃）を折れ線グラフで表したものである。また図2は，各月の日ごとの平均気温の分布を箱ひげ図で表したものである。(*)

図1

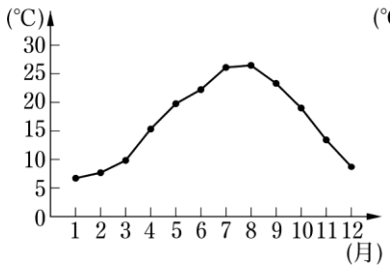
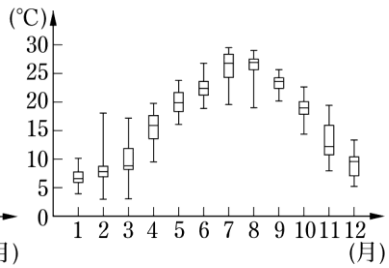
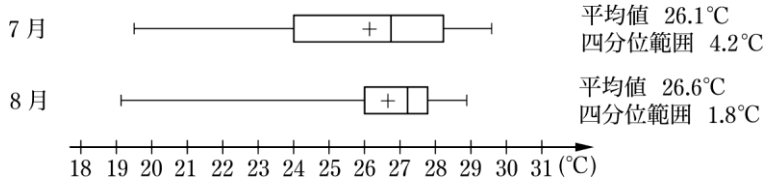



図2




平均値のみでは，各月の日ごとの平均気温の分布のようすを比較することができないが，箱ひげ図を用いることで分布のようすを比較することが可能である。

たとえば，7月と8月の箱ひげ図は下の図のようになる。




この図より，7月と8月の平均値や範囲はあまり変わらないが，8月の方が四分位範囲が小さい。ゆえに，8月の方が，日ごとの平均気温の散らばりが小さいことがわかる。

このように，箱ひげ図は複数のデータの分布のようすを比較するときに有効である。

**問 11** 次の文章は上の図2の1月と2月の日ごとの平均気温の分布について述べたものである。この文章が適切であるかどうかを答えよ。

「1月と2月において，1日の平均気温が5℃以上10℃以下であった日はそれぞれ月の半分以上ある。」

➡ p.170 問題1

---

(*)箱ひげ図は90°回転したものも用いられる。

## 5 分散と標準偏差

データの散らばり具合いを数値で表すために，データの個々の値と平均値の差に着目してみよう。

データの $n$ 個の値を $x_1,\ x_2,\ \cdots,\ x_n$ とし，その平均値を $\overline{x}$ とするとき，各値と平均値の差 $x_1-\overline{x},\ x_2-\overline{x},\ \cdots,\ x_n-\overline{x}$ をそれぞれ平均値からの**偏差**という。偏差の平均値を計算すると

$$\frac{1}{n}\{(x_1-\overline{x})+(x_2-\overline{x})+\cdots+(x_n-\overline{x})\}=\frac{1}{n}\{(x_1+x_2+\cdots+x_n)-n\overline{x}\}$$

$$=\frac{1}{n}(x_1+x_2+\cdots+x_n)-\overline{x}=\overline{x}-\overline{x}=0$$

となり，偏差の平均値ではデータの散らばり具合いを表すことはできない。

そこで，偏差を 2 乗した値 $(x_i-\overline{x})^2$ を考える。これらの値はすべて 0 以上であり，データの値 $x_i$ が平均値 $\overline{x}$ から離れているほど大きくなる。

したがって，偏差を 2 乗し，その平均値を求めると，データの散らばり具合いを表すことができる。この値を**分散**といい，$\boldsymbol{s^2}$ で表す。

分散は，計算の過程で数値を 2 乗するため単位が変わる。そこで，分散の正の平方根をとり，単位をデータの値に合わせる。この値を**標準偏差**といい，$\boldsymbol{s}$ で表す。(*)

一般に，分散や標準偏差が大きくなるほど，データの各値が平均値から離れており，散らばりが大きい。

| 分散と標準偏差 |
| --- |
| 分散　$s^2=\frac{1}{n}\{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2\}$<br>標準偏差　$\boldsymbol{s=\sqrt{\frac{1}{n}\{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2\}}}$<br>ただし，$\overline{x}$ は平均値 |

(*) $s$ は標準偏差を意味する standard deviation の頭文字である。

**例 8** 次のデータは，A 社と B 社の音楽プレイヤーを充電してからの連続使用時間を 5 回ずつ調べた結果である。

| | | | | | |
|---|---|---|---|---|---|
| A 社 | 24 | 28 | 26 | 22 | 25(単位　時間) |
| B 社 | 27 | 28 | 25 | 21 | 29(単位　時間) |

このとき，A 社の音楽プレイヤーの連続使用時間の標準偏差を求めてみよう。

連続使用時間を $x$ として，その平均値 $\overline{x}$ を求めると

$$\overline{x}=\frac{1}{5}(24+28+26+22+25)=\frac{1}{5}\cdot 125=25\ (\text{時間})$$

連続使用時間の各値の偏差は，下の表のようになる。

| 連続使用時間　$x$ | 24 | 28 | 26 | 22 | 25 |
|---|---|---|---|---|---|
| 連続使用時間の偏差 $x-\overline{x}$ | $-1$ | 3 | 1 | $-3$ | 0 |

上の表より，分散 $s^2$ を求めると

$$s^2=\frac{1}{5}\{(-1)^2+3^2+1^2+(-3)^2+0^2\}=\frac{1}{5}\cdot 20=4$$

よって，標準偏差 $s$ は　　$s=\sqrt{4}=2\ (\text{時間})$

**問 12** 例 8 における B 社の音楽プレイヤーの連続使用時間の標準偏差を小数第 3 位を四捨五入して，小数第 2 位まで求めよ。ただし，$\sqrt{2}=1.414$ とする。

次に，度数分布から標準偏差を求める方法を考えてみよう。変量 $x$ の平均値を $\overline{x}$ とすると，$x$ の分散 $s^2$ は次の式で計算する。

| 階級値 | 度数 |
|---|---|
| $x_1$ | $f_1$ |
| $x_2$ | $f_2$ |
| ⋮ | ⋮ |
| $x_r$ | $f_r$ |
| 計 | $n$ |

$$s^2=\frac{1}{n}\{(x_1-\overline{x})^2f_1+(x_2-\overline{x})^2f_2+\cdots+(x_r-\overline{x})^2f_r\}$$

よって，$x$ の標準偏差 $s$ は次の式で表される。

$$\boldsymbol{s=\sqrt{\frac{1}{n}\{(x_1-\overline{x})^2f_1+(x_2-\overline{x})^2f_2+\cdots+(x_r-\overline{x})^2f_r\}}}$$

**例題1 標準偏差**

右の度数分布表は, ある高校のバスケットボール部の部員 10 人でフリースローを 5 回ずつ行ったゲームの結果である。このゲームの得点の標準偏差を小数第 3 位を四捨五入して，小数第 2 位まで求めよ。

| 得点 $x$ | 度数 $f$ |
|---|---|
| 0 | 1 |
| 1 | 0 |
| 2 | 3 |
| 3 | 1 |
| 4 | 4 |
| 5 | 1 |
| 計 | 10 |

**解**

| 得点 $x$ | 度数 $f$ | $xf$ | $x-\overline{x}$ | $(x-\overline{x})^2$ | $(x-\overline{x})^2f$ |
|---|---|---|---|---|---|
| 0 | 1 | 0 | $-3$ | 9 | 9 |
| 1 | 0 | 0 | $-2$ | 4 | 0 |
| 2 | 3 | 6 | $-1$ | 1 | 3 |
| 3 | 1 | 3 | 0 | 0 | 0 |
| 4 | 4 | 16 | 1 | 1 | 4 |
| 5 | 1 | 5 | 2 | 4 | 4 |
| 計 | 10 | 30 | | | 20 |

$xf$ の値は上の表のようになるから，このゲームの得点の平均値 $\overline{x}$ は

$$\overline{x}=\frac{1}{10}\cdot 30=3\ (\text{点})$$

さらに，$x-\overline{x}$, $(x-\overline{x})^2$, $(x-\overline{x})^2f$ の値は上の表のようになるから

得点の分散 $s^2$ は　　$s^2=\frac{1}{10}\cdot 20=2$

よって，標準偏差 $s$ は　　$s=\sqrt{2}=1.414\cdots \fallingdotseq 1.41\ (\text{点})$

**問 13** 下の表は，あるクラス 40 人に数学の小テストを行った結果である。
このクラスの小テストの得点の標準偏差を小数第 3 位を四捨五入して, 小数第 2 位まで求めよ。ただし，$\sqrt{3}=1.732$ とする。

| 得点 $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 計 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 人数 | 2 | 4 | 7 | 8 | 4 | 6 | 2 | 1 | 2 | 3 | 1 | 40 |

**分散と平均値の関係式**

166 ページで示した分散 $s^2$ を表す式は，次のように変形できる。

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

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

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

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

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

したがって，データの値を 2 乗した値の平均値から，平均値の 2 乗を引いた値が分散に等しいことがわかる。

一般に，変量 $x$ の分散について，次のことが成り立つ。

$$(\boldsymbol{x}\text{ の分散}) = (\boldsymbol{x^2}\text{ の平均値}) - (\boldsymbol{x}\text{ の平均値})^2$$

**例 9** 下の表は，高校 1 年生の A さんが毎朝通学に利用しているバスの乗車時間を 6 日間調べた結果である。

| バスの乗車時間 $x$（分） | 9 | 13 | 10 | 9 | 10 | 12 |
|---|---|---|---|---|---|---|

上の計算式を用いて，バスの乗車時間の分散を求めてみよう。

$x$ の平均値 $\overline{x}$ と $x^2$ の平均値 $\overline{x^2}$ は，次のようになる。

$$\overline{x} = \frac{1}{6}(9 + 13 + 10 + 9 + 10 + 12) = \frac{1}{6} \cdot 63 = \frac{21}{2}$$

$$\overline{x^2} = \frac{1}{6}(9^2 + 13^2 + 10^2 + 9^2 + 10^2 + 12^2) = \frac{1}{6} \cdot 675 = \frac{225}{2}$$

よって，分散 $s^2$ は　　$s^2 = \frac{225}{2} - \left(\frac{21}{2}\right)^2 = \frac{9}{4} = 2.25$

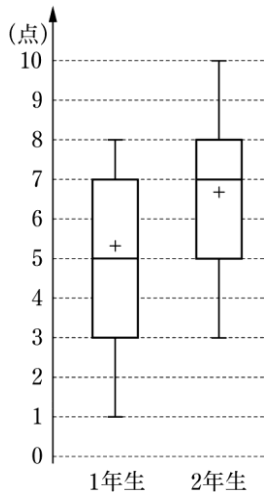
**問 14** 下の表は，A さんの 1 日の学習時間を 5 日間記録したものである。

| 1 日の学習時間 $x$（時間） | 1 | 2 | 3 | 1 | 3 |
|---|---|---|---|---|---|

このとき，1 日の学習時間の分散を求めよ。

**問題**

**1** 右の図は，ある学校の1年生と2年生それぞれ20人に対して，10点満点の漢字テストを実施したときの得点の箱ひげ図である。このとき，次の①～④のうち，右の図から必ずしも正しいとは限らないものをすべて選べ。




① 1年生の得点のデータの方が2年生の得点のデータより四分位範囲が大きい。

② どちらの学年にも，平均点以下の生徒が10人以上いる。

③ どちらの学年にも，得点が5点以上の生徒が10人以上いる。

④ どちらの学年にも，得点がちょうど8点の生徒がいる。

**2** くじを20回引いて，当たった回数だけ得点できるゲームがある。右の表は，ある学校の生徒5人がこのゲームを行ったときの得点を記録したものである。ただし，生徒4の得点は5人の得点の平均値以下であった。このとき，次の問に答えよ。 参考 P. 179

| | 得点 |
|---|---|
| 生徒1 | 8 |
| 生徒2 | 14 |
| 生徒3 | 10 |
| 生徒4 | $a$ |
| 生徒5 | $18-a$ |
| 平均値 | $m$ |
| 分散 | 6 |

(1) 5人の得点の平均値 $m$ を求めよ。

(2) 表中の $a$ の値を求めよ。

(3) 右の表の得点に対して，全員に10点を加えたとき，分散の値はどのように変化するか。①～③のうちから適するものを選べ。

① 大きくなる ② 変わらない ③ 小さくなる