数学Ⅰ　Standard　p.161

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。』
　　フローレンス・ナイチンゲール
　　（1820年～1910年）

1節 データの整理と分析

❶ データの整理

資料を集め，それを整理して，いろいろなことを調べることはよく行われる。次の資料は，ある年の9月の大阪の最高気温を日付順に横に並べたものである。

【データ1 9月の大阪の最高気温（単位 ℃)】

31.5	33.4	31.2	32.9	34.0	34.5	33.2	31.3	27.8	28.0
30.8	25.3	28.1	27.6	22.6	28.2	29.5	27.6	28.5	28.4
30.0	27.7	29.1	31.5	31.1	30.8	31.5	27.7	26.0	22.0

このような資料を**データ**という。また，気温，湿度，降水量のように，ある特性を数量的に表すものを**変量**という。

度数分布表

データについて調べるとき，上のデータのようにただ個々の値を並べただけでは，全体の特徴がつかみにくい。このようなときには，中学校で学んだように，表やグラフなどを用いてデータの特徴を把握するとよい。

右の表は，22℃から36℃までを2℃ごとの区間に分けて，データ1を整理したものである。このような区間を**階級**といい，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，各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。

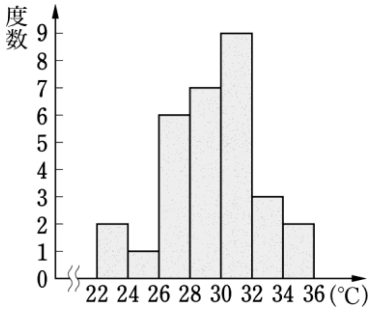
たとえば，右の度数分布表において，「26℃以上28℃未満」の階級の幅は2℃，階級値は27℃，階級の度数は6である。

気温の 階級 ℃以上℃未満	度数
22～ 24	2
24～ 26	1
26～28	6
28～30	7
30～32	9
32～34	3
34～36	2
計	30

度数分布のグラフ

度数分布は，**ヒストグラム**とよばれるグラフで表すことが多い。ヒストグラムは，階級の幅が一定のとき，長方形の高さが度数を表すようにかく。

前のページのデータ１の度数分布のヒストグラムをかくと，右のようになる。



問 1 次のデータは，ある年の９月の東京の最高気温を気温の低い順に横に並べたものである。22℃以上 24℃未満という階級から，幅を一定にとって順次階級をつくり，このデータの度数分布表を作成せよ。また，そのヒストグラムをかけ。

【データ２　９月の東京の最高気温（単位　℃)】

23.2	23.3	23.4	23.7	23.8	23.9	24.0	24.4	24.5	25.1
25.7	25.8	26.2	26.4	26.5	26.5	26.9	26.9	27.6	27.8
28.1	28.3	28.5	28.5	28.8	28.8	28.9	29.0	29.3	31.5

相対度数

度数分布表において，各階級の度数を度数の合計で割った値を**相対度数**という。

$$\text{相対度数} = \frac{\text{度数}}{\text{度数の合計}}$$

右の表で，たとえば，26℃以上 28℃未満の階級の相対度数は，次のようになる。

$$\frac{6}{30} = 0.2$$

度数分布表には，必要に応じて相対度数などを並べて記入することもある。

気温の階級 ℃以上 ℃未満	度数	相対度数
22 ~ 24	2	0.07
24 ~ 26	1	0.03
26 ~ 28	6	0.20
28 ~ 30	7	0.23
30 ~ 32	9	0.30
32 ~ 34	3	0.10
34 ~ 36	2	0.07
計	30	1.00

問 2 問１で作成した度数分布表において，各階級の相対度数を求めよ。

❷ データの代表値

データの特徴を1つの数値で表すことによって，データの傾向を調べてみよう。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

変量 x の n 個の値 x_1，x_2，…，x_n からなるデータがあるとき，これらの総和を n で割った値をデータの**平均値**といい，記号 $\overline{x}$ で表す。

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

例 1 162 ページのデータ 1 の平均値は

$$\frac{31.5+33.4+\cdots+22.0}{30}=\frac{881.8}{30}\fallingdotseq 29.4\ (℃)$$

度数分布から求めた平均値

度数分布が与えられたとき，階級値 x と度数 f の積 xf を求めて，その総和を度数の合計で割ったものを，度数分布から求めた平均値という。

例 2 データ 1 の平均値を度数分布から求めてみよう。

度数分布表に積 xf の欄をつけ加えると右のようになるから，求める平均値は

$$\frac{884}{30}\fallingdotseq 29.5\ (℃)$$

この値は，例 1 で計算した平均値に近い値となっている。

階級 ℃ 以上 ℃ 未満	階級値 x	度数 f	xf
22 ~ 24	23	2	46
24 ~ 26	25	1	25
26 ~ 28	27	6	162
28 ~ 30	29	7	203
30 ~ 32	31	9	279
32 ~ 34	33	3	99
34 ~ 36	35	2	70
計		30	884

問 3 163 ページの問 1 で作成した度数分布表から，平均値を小数第 2 位を四捨五入して求めよ。

中央値，最頻値

データのすべての値を小さい順に並べたとき，中央の位置にある数値を**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，すなわち $2n$ 個のときは，第 n 番目と第 $n+1$ 番目の数値の平均値を中央値とする。

例 3 162 ページのデータ 1 の中央値を求めてみよう。

データの値を小さい順に並べかえると次のようになる。

22.0 22.6 25.3 26.0 27.6 27.6 27.7 27.7 27.8 28.0
28.1 28.2 28.4 28.5 29.1 29.5 30.0 30.8 30.8 31.1
31.2 31.3 31.5 31.5 31.5 32.9 33.2 33.4 34.0 34.5

データの値の個数は 30 個であるから 15 番目の 29.1 と 16 番目の 29.5 の平均値が求める中央値である。

よって，データ 1 の中央値は $\frac{29.1+29.5}{2} = 29.3$ (℃)

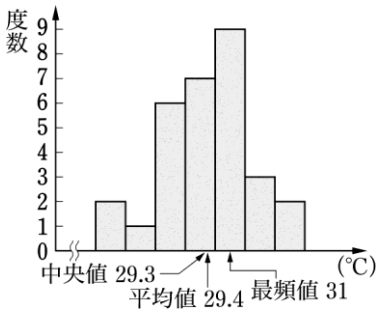
度数分布表で，度数がもっとも多い階級の階級値を**最頻値**または**モード**という。すなわち，最頻値はヒストグラムにおいて，もっとも高い長方形の階級値である。*

例 4 162 ページの度数分布表からデータ 1 の最頻値を求めてみよう。度数がもっとも多い階級は「30℃以上 32℃未満」であるから，その階級値 31℃が最頻値である。

問 4 163 ページの問 1 のデータ 2 の中央値を求めよ。また，問 1 で作成した度数分布表から最頻値を求めよ。

p.173 Training 1

162 ページのデータ 1 において，例 1，例 3，例 4 で求めた平均値，中央値，最頻値をヒストグラム上に表すと右の図のようになる。



*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。

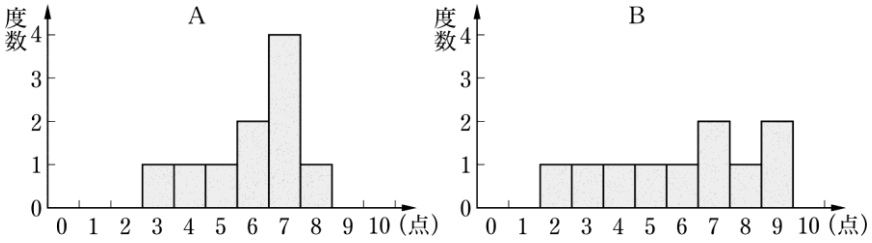
❸ データの散らばり

A，B の 2 人で 10 点満点の的当てゲームを 10 回行い，成績は右の表のようになった。

回	1	2	3	4	5	6	7	8	9	10
Aの得点	7	3	6	4	8	7	7	5	7	6
Bの得点	9	7	2	8	6	3	4	7	9	5

このとき，A，B ともに平均値は 6 点，中央値は 6.5 点である。

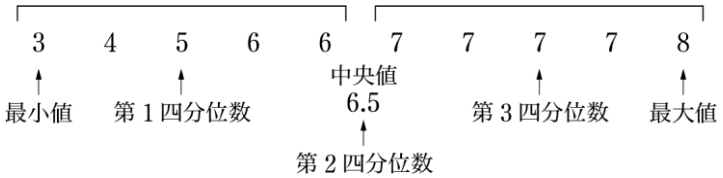
しかし，下のヒストグラムから，2 人の得点の分布には散らばり具合いに違いがみられる。ここでは，データの分布の特徴について，散らばり具合いを含めて考えてみよう。



四分位数

中央値をもとに，データの分布を表すための数値を考えてみよう。

上の的当てゲームにおける A の得点を小さい順に並べかえ，中央値を境にして 2 つの部分に分ける。



最小値を含む方の 5 個のデータの中央値 5 点を**第 1 四分位数**という。

中央値 6.5 点を**第 2 四分位数**という。

最大値を含む方の 5 個のデータの中央値 7 点を**第 3 四分位数**という。

これらを合わせて**四分位数**という。*

---

*第 1 四分位数，第 2 四分位数，第 3 四分位数をそれぞれ Q_1，Q_2，Q_3 と書くこともある。

一般に，四分位数は，データの値の個数が偶数 $(2n)$，奇数 $(2n+1)$ のいずれのときも，次のように求める。

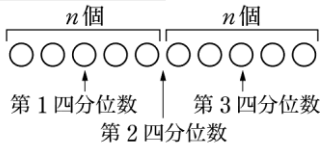
データの値を小さい順に並べかえて，右の図のように，中央値を境にして 2 つの部分に分ける。

① 最小値を含む方の n 個のデータの中央値が第 1 四分位数である。

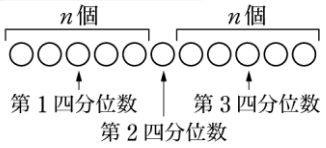
② 中央値が第 2 四分位数である。

③ 最大値を含む方の n 個のデータの中央値が第 3 四分位数である。

偶数 $(2n)$ のとき



奇数 $(2n+1)$ のとき



例 5 前のページの的当てゲームにおける B の得点の四分位数を求めてみよう。B の得点を小さい順に並べかえると

2　3　4　5　6　7　7　8　9　9

第 1 四分位数は，2，3，4，5，6 の中央値より　4 点

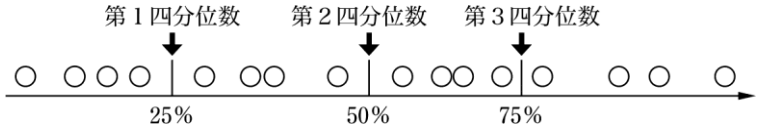
第 2 四分位数は，データの中央値より　$\frac{6+7}{2}=6.5$ (点)

第 3 四分位数は，7，7，8，9，9 の中央値より　8 点

問 5 次のデータはあるクラスの 1 班，2 班の 1 か月の読書時間である。1 班，2 班の四分位数をそれぞれ求めよ。

1 班 20 人（単位　時間）	2 班 15 人（単位　時間）
3 10 7 14 5 9 15 0 13 18	20 6 0 14 16 23 1 4 5 0
0 8 11 10 15 19 6 23 9 5	18 13 21 0 9

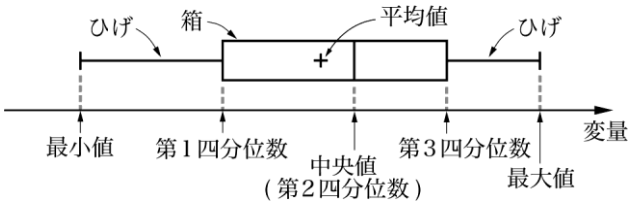
第 1 四分位数は，データの値を小さい順に並べたとき，値の小さい方から 25% の位置を示す値である。同様に，第 2 四分位数，第 3 四分位数は，値の小さい方から 50%，75% の位置を示す値である。



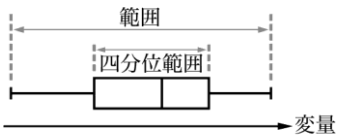
箱ひげ図

データの分布を最小値，第1四分位数，中央値（第2四分位数），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。

5数要約を用いてデータの分布を表すには，**箱ひげ図**を用いる。箱ひげ図とは，次の図のように，5数要約を箱と線（ひげ）を用いて1つの図に表したものである。さらに平均値の位置を表すこともある。

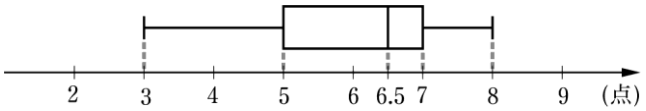


データの最大値から最小値を引いた値を**範囲**または**レンジ**という。また，第3四分位数から第1四分位数を引いた値を**四分位範囲**といい，四分位範囲を2で割った値を**四分位偏差**という。箱ひげ図では，ひげを含めた全長が範囲を表し，箱の横の長さが四分位範囲を表す。



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

例6 166ページの的当てゲームにおけるAの得点の箱ひげ図をかいてみよう。



Aの得点の範囲は　$8-3=5$（点）

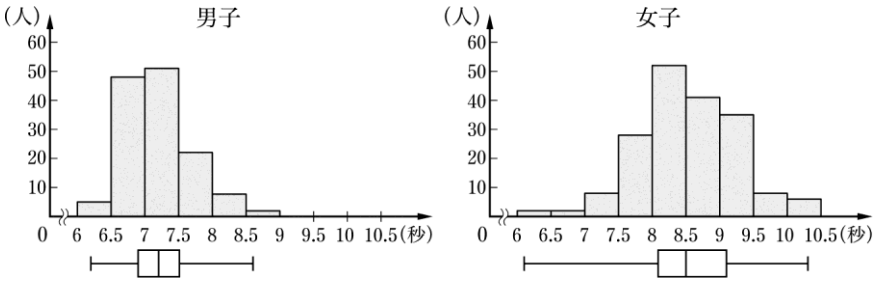
四分位範囲は　$7-5=2$（点）

四分位偏差は　$\frac{2}{2}=1$（点）

問6 166ページの的当てゲームにおけるBの得点の箱ひげ図をかけ。また，Bの得点の範囲，四分位範囲，四分位偏差を求めよ。

ヒストグラムと箱ひげ図

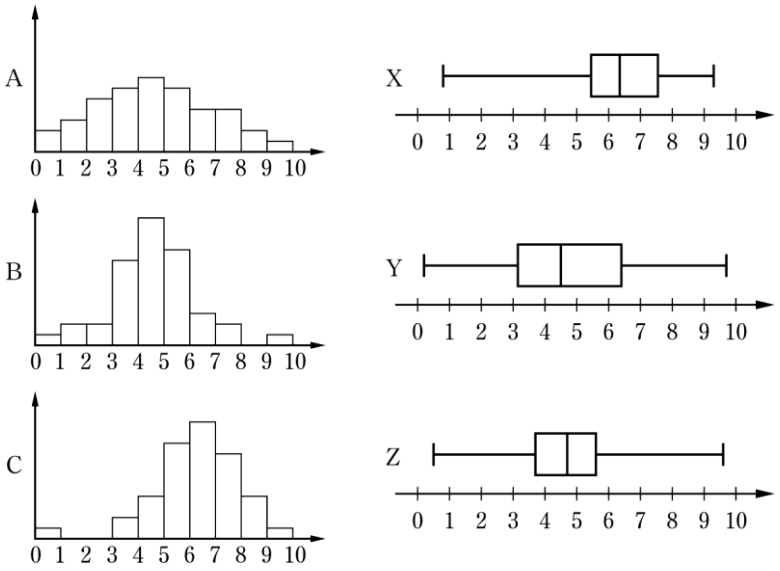
下のヒストグラムと箱ひげ図は，ある高校の1年生の体力テストにおける50m走の結果を，男子と女子に分けて表したものである。



上の図では，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。一般に，分布が1つの山の形をしたヒストグラムとなる場合，箱ひげ図をみることによってヒストグラムのおおよその形を知ることができる。

問 7 次のA，B，Cのヒストグラムについて，それぞれに対応する箱ひげ図として適切なものをX，Y，Zの中から選べ。

p.173 Training 2



箱ひげ図とデータの散らばり

次の図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を折れ線グラフで表し，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

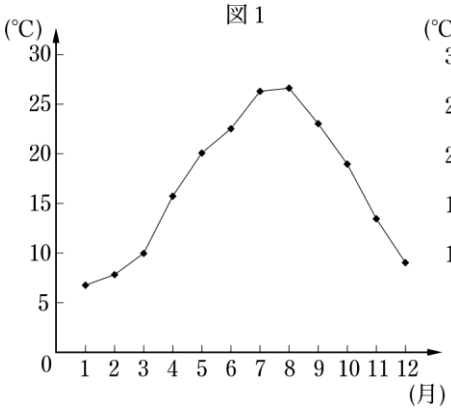


図2

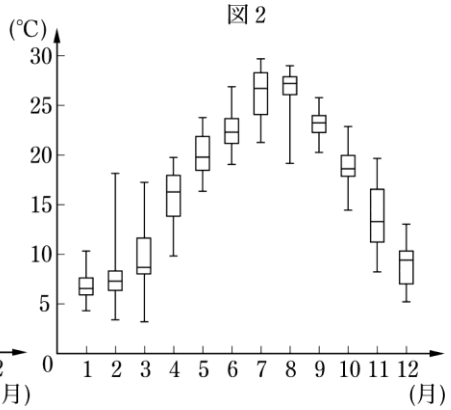
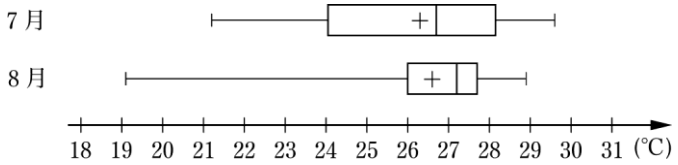


図1のように,平均値のみでは各月の日ごとの平均気温の分布のようすを比較することはできないが，図2のように箱ひげ図を用いることで，分布のようすを，散らばり具合いを含めて比較することができる。

たとえば，7月と8月の箱ひげ図は下の図のようになる。



この箱ひげ図より，7月と8月の平均気温の平均値はあまり変わらないが，箱の横の長さやひげの長さを比較することにより，データの散らばり具合いの違いがわかる。

このように，箱ひげ図は複数のデータの分布を比較するのに適している。

問8 167ページの問5の1班, 2班について，それぞれ箱ひげ図をかき，分布を比較せよ。

p.182 Level Up 1

---

*箱ひげ図は90°回転したものも用いられる。

分散と標準偏差

データの散らばり具合いを，個々の値と平均値との差を用いて表すことを考えてみよう。

データの個々の値が x_1，x_2，…，x_n であり，その平均値を $\overline{x}$ とするとき

$$x_1-\overline{x},\quad x_2-\overline{x},\quad \cdots,\quad x_n-\overline{x}$$

をそれぞれの値の**平均値からの偏差**あるいは単に**偏差**という。

偏差の総和を計算すると

$$\begin{aligned}&(x_1-\overline{x})+(x_2-\overline{x})+\cdots+(x_n-\overline{x})\\&=(x_1+x_2+\cdots+x_n)-n\overline{x}\\&=(x_1+x_2+\cdots+x_n)-n\cdot\frac{1}{n}(x_1+x_2+\cdots+x_n)=0\end{aligned}$$

となるから，偏差の平均値は 0 になる。したがって，偏差の平均値ではデータの散らばり具合いを表すことができない。

次に，偏差を 2 乗した値

$$(x_1-\overline{x})^2,\quad (x_2-\overline{x})^2,\quad \cdots,\quad (x_n-\overline{x})^2$$

を考えると，これらの値はすべて 0 以上で，データの値が平均値 $\overline{x}$ から離れているほど大きくなる。

したがって，偏差の 2 乗の平均値

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

を求めれば，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，単位がデータの値の単位とは異なる。そこで，単位をそろえるために，分散の正の平方根を考える。

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

これを**標準偏差**といい，$\boldsymbol{s}$ で表す。分散と同じように，標準偏差もデータの散らばり具合いを表す。

分散と標準偏差

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

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

ただし，$\overline{x}$ は平均値

例 7 166 ページの的当てゲームにおける A の得点の分散 ${s_A}^2$，標準偏差 s_A を求めてみよう。

A の得点の平均値は 6 点であるから，偏差は次の表のようになる。

回	1	2	3	4	5	6	7	8	9	10
得点	7	3	6	4	8	7	7	5	7	6
偏差	1	−3	0	−2	2	1	1	−1	1	0

よって

$${s_A}^2 = \frac{1}{10}\{1^2 + (-3)^2 + 0^2 + (-2)^2 + 2^2 + 1^2 + 1^2 + (-1)^2 + 1^2 + 0^2\}$$

$$= 2.2$$

$$s_A = \sqrt{2.2} \fallingdotseq 1.48\ (\text{点})$$ ―― 電卓を用いて求めるとよい

166 ページの的当てゲームにおける B の得点の分散 ${s_B}^2$，標準偏差 s_B を，例 7 と同様に求めると

$${s_B}^2 = 5.4$$

$$s_B = \sqrt{5.4} \fallingdotseq 2.32\ (\text{点})$$

となる。したがって，分散，標準偏差の値は A より B の方が大きいから，得点の散らばり具合いは B の方が大きいことがわかる。

問 9 右の表はある生徒の 5 回のテストの得点である。得点の分散と標準偏差を求めよ。

回	1	2	3	4	5
得点	82	76	63	80	64

p.173 Training 3　p.182 Level Up 2

Training

1 右の表は，50 人の生徒が受けたあるテストの結果を，度数分布表にまとめたものである。それぞれの階級の点数をとった生徒の数を度数とする。このとき，次の問に答えよ。

(1) 表の空欄をうめて度数分布表を完成させ，最頻値を求めよ。

(2) テストの点数の平均値を度数分布表から求めよ。 p.165

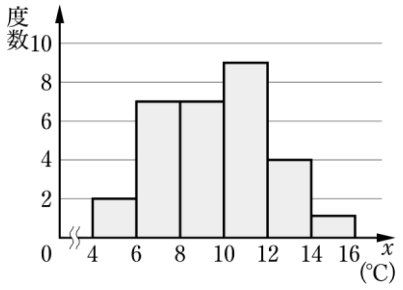
点数の階級 点以上　点未満	度数	相対度数
0 ~ 10		0.12
10 ~ 20	8	0.16
20 ~ 30		
30 ~ 40		0.28
40 ~ 50	12	0.24
計	50	

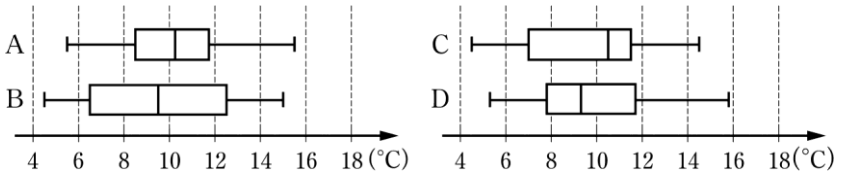
2 ある都市における 1 日の平均気温を調べた。次の問に答えよ。

(1) 次のデータは 3 月 1 日から 10 日までの 1 日の平均気温を順に並べたものである。このデータの箱ひげ図をかけ。

2.6　1.8　1.3　2.5　5.8　5.8　5.3　6.4　6.2　6.9　（単位℃）

(2) 右の図は 11 月の 30 日間における 1 日の平均気温のデータのヒストグラムである。このデータを箱ひげ図にまとめると，次の A ~ D のいずれかになった。このデータの箱ひげ図は A ~ D のどれか。





p.169

3 A，B の 2 人が 10 点満点の小テストを 5 回受け，右の表のデータを得た。A，B それぞれの得点の分散を求めよ。 p.172

回	1	2	3	4	5
Aの得点	7	5	8	6	4
Bの得点	2	7	10	5	6

参考　分散の計算

5 個の値 x_1，x_2，x_3，x_4，x_5 からなるデータがある。このデータの平均値を $\overline{x}$ とすると

$$\frac{1}{5}(x_1+x_2+x_3+x_4+x_5)=\overline{x}$$

が成り立つ。このとき，データの分散を s^2 とすると

$$s^2=\frac{1}{5}\{(x_1-\overline{x})^2+(x_2-\overline{x})^2+(x_3-\overline{x})^2+(x_4-\overline{x})^2+(x_5-\overline{x})^2\}$$

$$=\frac{1}{5}\{({x_1}^2+{x_2}^2+{x_3}^2+{x_4}^2+{x_5}^2)-2\overline{x}(x_1+x_2+x_3+x_4+x_5)+5(\overline{x})^2\}$$

$$=\frac{1}{5}({x_1}^2+{x_2}^2+{x_3}^2+{x_4}^2+{x_5}^2)-2\overline{x}\cdot\frac{1}{5}(x_1+x_2+x_3+x_4+x_5)+(\overline{x})^2$$

$$=\frac{1}{5}({x_1}^2+{x_2}^2+{x_3}^2+{x_4}^2+{x_5}^2)-2(\overline{x})^2+(\overline{x})^2$$

$$=\frac{1}{5}({x_1}^2+{x_2}^2+{x_3}^2+{x_4}^2+{x_5}^2)-(\overline{x})^2$$

となり，データの値を 2 乗した値の平均値から，データの平均値の 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 の分散は 172 ページの求め方のほかに

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

の式を用いても求めることができる。

また，変量 x の標準偏差についても，次のことが成り立つ。

$$\boldsymbol{(x\text{の標準偏差})=\sqrt{(x^2\text{の平均値})-(x\text{の平均値})^2}}$$

例 1　上の分散の計算式を用いて，166 ページの的当てゲームにおける A の得点の分散を求めてみよう。

$${s_\mathrm{A}}^2=\frac{1}{10}(7^2+3^2+6^2+4^2+8^2+7^2+7^2+5^2+7^2+6^2)-6^2$$

$$=\frac{1}{10}\times 382-36=2.2$$

問 1　上の計算式を用いて，166 ページの的当てゲームにおける B の得点の分散を求めよ。