数学Ⅰ・A 基礎問題精講［四訂増補版］

上園信武著

第 8 章　データの分析

旺文社

第8章 データの分析

130 度数分布表とヒストグラム

次のデータは，あるクラス 30 人に行った 100 点満点の数学のテストの得点の結果である．

64，32，81，59，47，53，55，42，77，78，89，63，33，68，61，
59，48，76，63，77，83，95，56，62，68，76，66，70，44，65

(1)　階級の幅を 10 点として，度数分布表をつくれ．ただし，階級は 30 点から区切り始めるものとする．
(2)　(1)の度数分布表をもとにして，ヒストグラムをかけ．

精講　テストの点数や，人の身長・体重，あるいは 50 m 走のような運動の記録のように，ある特性を表す数量を**変量**といい，ある変量の測定値を集めたものを**データ**といいます．

このデータをいくつかの幅で区切って階級を定め，各階級に属するデータの個数を対応させた表を**度数分布表**といい，各階級の中央の値を**階級値**といいます(たとえば，この度数分布表では，階級値は小さい方から，35，45，55，65，75，85，95 である)．

また，度数分布表を柱状のグラフで表したものを**ヒストグラム**といいます．(このようにグラフにすることによって，データを視覚的にとらえることができる)

参考　ヒストグラム(histogram)という用語は，histo＋gram で，histo が「織り物」，gram が「表現されたもの(＝文書，図表)」というギリシャ語から来ています．さしずめ，「データ(数値)を織り込んだ図」という意味になるのでしょう．また，これとは逆に日本語になっている数学用語で漢字からは意味が想像つかないものもあります．皆さん方が中学で学んだ「座標」などもそうでしょう．これは，英語で「coordinate」といいますが，ファッションの世界で「上下のコーディネートが良くない」などと使いますね．

なぞかけ風にいうと「座標とかけて，ファッションと解く．そのココロはどちらも組み合わせます」という感じでしょうか？　座標は数字を，ファッションは洋服を組み合わせるわけです．

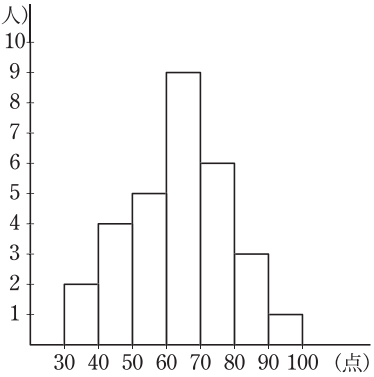
こんな角度から数学をながめるのもおもしろいかもしれません．普通の英和辞典などでも，[数] などの記号付きで訳が載っています．興味ある人は，「数学英和・和英辞典」などを入手する手もありでしょう．

解　答

(1)

階　級 (点)	度数
30 以上 ～ 40 未満	2
40 ～ 50	4
50 ～ 60	5
60 ～ 70	9
70 ～ 80	6
80 ～ 90	3
90 ～ 100	1
計	30

(2)



ポイント　度数分布表をつくるときは，まず階級の幅を決め，それぞれの階級に属するデータを数え上げて表にする

演習問題 130

次のデータは，ある弁当屋のある月の 1 日毎の弁当の売り上げ個数である．

127，116，182，188，171，133，139，162，179，154，128，
144，166，150，155，141，156，148，147，159，137，123，
161，123，176，125，147，113，191，186

(1)　階級の幅を 10 個として，度数分布表をつくれ．ただし，階級は 110 個から区切り始めるものとする．

(2)　(1)の度数分布表をもとにして，ヒストグラムをかけ．

131 データの代表値（平均値・メジアン・モード）

右表は100点満点の数学のテストの結果を度数分布表にしたものである．この表をもとにして，以下の問いに答えよ．

(1)　最頻値を求めよ．

(2)　階級値を用いて平均値を求めよ．

(3)　得点が40点以上60点未満の階級に含まれる8人の得点は，以下のようになっていた．

41，56，50，42，51，59，41，50

このとき，この階級における中央値と平均値を求めよ．

得　点（点）	度数
0 以上 ～ 20 未満	2
20 ～ 40	11
40 ～ 60	8
60 ～ 80	15
80 ～ 100	4
計	40

精講

データが度数分布表の形で表されているとき，そのデータの特徴を示す値を**代表値**といいます．代表値として我々が日頃耳にするのは，最高，最低，平均などですが，数学では，**平均値**，**最頻値**（モード），**中央値**（メジアン）の3つがよく用いられます．まず，それぞれの定義をはっきりさせておきましょう．

①**平均値**：変量xのデータの値が，x_1，x_2，x_3，…，x_nのとき，平均値$\overline{x}$は，

$$\overline{x}=\frac{1}{n}(x_1+x_2+\cdots+x_n)$$ で表される．

この問題のように個々のデータがなく，度数分布表でデータが与えられているときは，個々のデータはすべて階級値（⇨130）とみなして，平均値を求める（⇨**解答**(2)）．

②**最頻値**（モード）：データにおいて最も多い値．度数分布表では，最も度数の多い階級の階級値．

③**中央値**（メジアン）：データを大きい順（または小さい順）に並べたとき，その中央にくる値．データの個数が偶数のときは，中央の2つの値の平均．

解　答

(1)　データの最も度数の多い階級は 60 点以上 80 点未満だから，最頻値（モード）は，この階級の階級値で　$\dfrac{60+80}{2}=\mathbf{70}$ **(点)**

(2)　各階級の階級値は，小さい順に 10 点，30 点，50 点，70 点，90 点で，それぞれに対応する度数は，2 人，11 人，8 人，15 人，4 人だから，平均値は，

$$\frac{1}{40}(10\times2+30\times11+50\times8+70\times15+90\times4)=\frac{216}{4}=\mathbf{54}\ \textbf{(点)}$$

(3)　データを小さい順に並べると，41，41，42，50，50，51，56，59

よって，中央値は，$\dfrac{50+50}{2}=\mathbf{50}$ **(点)**

平均値は，$\dfrac{1}{8}(41+41+42+50+50+51+56+59)=\dfrac{390}{8}=\mathbf{48.75}$ **(点)**

ポイント　代表値（平均値・最頻値・中央値）を求めるときは，定義にしたがって計算する

注　度数分布表でデータが与えられているときの平均値は，階級値を使っているので**正確とはいいきれません**．しかし，平均値に幅をもたせて，平均値がどんな範囲にあるかは調べることができます．このときは，階級値ではなく，階級のとりうる値を利用して計算します（⇨**演習問題 131**）．

演習問題 131

右表はあるクラスの 50 m 走の度数分布表である．

(1)　最頻値を求めよ．

(2)　階級値を用いて平均値を求めよ．

(3)　階級値を用いないで，平均値を求めたとき，平均値のとりうる値の範囲を求めよ．

タイム（秒）	度　数
6.0（以上）～ 6.5（未満）	2
6.5 ～ 7.0	2
7.0 ～ 7.5	6
7.5 ～ 8.0	8
8.0 ～ 8.5	2
計	20

132 四分位数

次のデータはA君，B君の数学のテストの得点である．

A君：64，32，81，59，47，53，55，42，77，78，89，63，33，68，61

B君：58，48，76，63，77，83，95，56，62，68，76，66，70，44，65

(1)　A君，B君のそれぞれのデータについて，四分位数，四分位範囲，四分位偏差を求めよ．

(2)　A君とB君のデータについて，四分位範囲を比べることによって，データの散らばり度合いを比較せよ．

精講　データの散らばりの度合いを比べる1つのものさしとして**四分位範囲**というものがあります．これを求めるためには，まず**四分位数**という数値を求める必要があります．これは次の手順で求めます．

① データを小さい順に並べる．このときの中央値（メジアン）を求める．
これを**第2四分位数**（$=Q_2$）といいます．

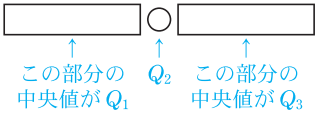
② Q_2 を境にしてデータを前半と後半に分け，前半部分の中央値を求める．これを**第1四分位数**（$=Q_1$）といいます．
次に，後半部分の中央値を求める．これを**第3四分位数**（$=Q_3$）といいます．

③ Q_3-Q_1 を**四分位範囲**，$\dfrac{Q_3-Q_1}{2}$ を**四分位偏差**といいます．

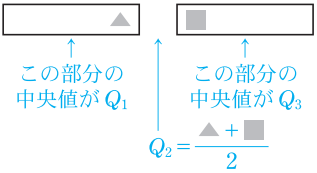
注　データの大きさが奇数のときは，Q_2 はデータの数値そのもので，データを2等分するときに Q_2 は含まず（⇨図Ⅰ），データの大きさが偶数のときは，Q_2 はデータそのものではなく，中央の2つの値の平均です（⇨図Ⅱ）．Q_1，Q_3 を求めるときも同じです．

また，「四分位数を求めよ」といわれたら，特に指定がない限り，第1四分位数，第2四分位数，第3四分位数をすべて答えます．

（図Ⅰ）データの大きさが奇数



（図Ⅱ）データの大きさが偶数



解　答

(1)　A君，B君のデータを小さい順に並べると次のようになる．

A君：32，33，42，47，53，55，59，61，63，64，68，77，78，81，89

B君：44，48，56，58，62，63，65，66，68，70，76，76，77，83，95

（A君について）

第2四分位数は**61点**，第1四分位数は**47点**，第3四分位数は**77点**より，

四分位範囲は　77−47＝**30**（**点**），四分位偏差は　30÷2＝**15**（**点**）

（B君について）

第2四分位数は**66点**，第1四分位数は**58点**，第3四分位数は**76点**より，

四分位範囲は　76−58＝**18**（**点**），四分位偏差は　18÷2＝**9**（**点**）

(2)　A君の四分位範囲の方がB君の四分位範囲より大きいので，A君の方がデータの散らばり度合いが大きい．

注　データを小さい順に並べたとき，次のようになっていると

x_1，x_2，x_3，x_4，x_5，x_6，x_7，x_8，x_9，x_{10}，x_{11}，x_{12}

$Q_2=\dfrac{x_6+x_7}{2}$，$Q_1=\dfrac{x_3+x_4}{2}$，$Q_3=\dfrac{x_9+x_{10}}{2}$　になります．

ポイント

四分位数の求め方

①データを小さい順に並べて

②中央値を考えて，第2四分位数を決定

③中央値より小さいデータの中央値を考えて第1四分位数を，中央値より大きいデータの中央値を考えて第3四分位数を求める

演習問題 132

次のデータは，A君，B君2人の生徒の10点満点のテストの結果である．

A君：1，2，2，5，6，10，5，6，2，1

B君：5，5，7，8，1，10，10，8，9，4

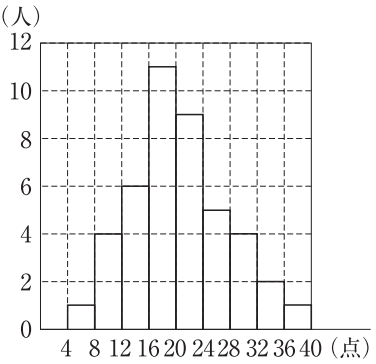
(1)　A君，B君それぞれについて，四分位数，四分位範囲を求めよ．

(2)　四分位範囲を比べることによって，データの散らばり度合いを比較せよ．

133 ヒストグラムと四分位数

ある高校3年生1クラスの生徒43人について，10点満点のテスト4回分の合計点のデータを取った．右の図は，このデータのヒストグラムである．

ただし，階級 a ～ b に属するとは得点が a 点以上 b 点未満であることを表し，テストの得点は整数値をとるものとする．



この43人のデータから，第1四分位数 Q_1，第2四分位数 Q_2（中央値），第3四分位数 Q_3 が含まれる階級の階級値を求めよ．

精講

132によると，43人のデータの場合の各四分位数はデータを小さい方から並べたとき，**11人目**，**22人目**，**33人目**になります．

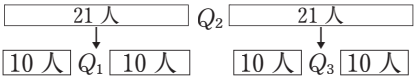
そこで，まず，ヒストグラムで小さい方から何人目かがわかるように順位の番号をつけておきます．（⇨**解答**の図を参照）

これで，各四分位数が属する階級がわかります．

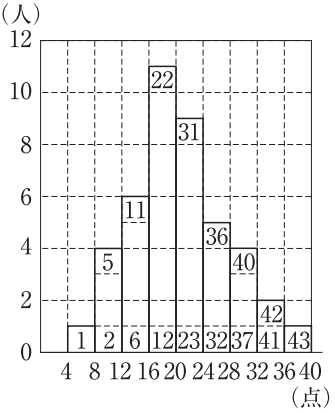
また，階級値は130によると，各階級の中央の値です．

解　答

右図のように，ヒストグラムに順位の番号をつけておく．43 人のデータを小さい順に並べたとき，Q_1，Q_2，Q_3 はそれぞれ，11 人目，22 人目，33 人目であるから，



Q_1，Q_2，Q_3 はそれぞれ，12 点～ 16 点，16 点～ 20 点，24 点～ 28 点の階級に含まれているので，求める階級値は，それぞれ，**14 点**，**18 点**，**26 点**である．

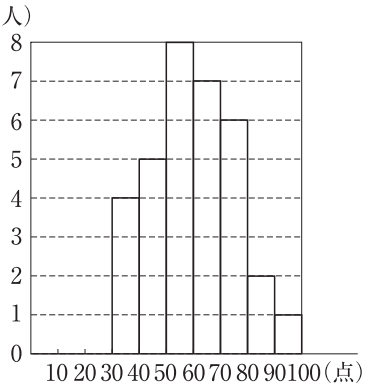


ポイント　データの各代表値の定義をしっかり覚えることが第一歩

演習問題 133

33 人の生徒に対して，100 点満点の試験をして，その結果をヒストグラムにすると，右の図のようになった．

このデータの第 1 四分位数，第 2 四分位数，第 3 四分位数が存在する階級の階級値をそれぞれ求めよ．



134 箱ひげ図

次の2つのデータは，JRのK線とI線の駅間の距離を並べたものである．ただし，単位はkmとする．

K線：4.0，1.5，0.8，3.5，1.9，2.8，1.1，2.7，2.2，3.0，2.1，5.1，2.1，5.1，5.2

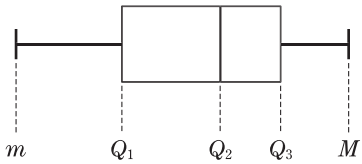
I線：2.2，1.3，1.4，2.5，1.8，3.1，3.4，2.6，4.2，4.6，2.9，2.8，2.7，2.3，4.3

(1) K線，I線それぞれについて箱ひげ図をかけ．

(2) 駅間距離の散らばり度合いはどちらが大きいといえるか．

精講

(1) 箱ひげ図とは，あるデータの最大値をM，最小値をm，第1四分位数をQ_1，第2四分位数をQ_2，第3四分位数をQ_3とするとき，これら**5つの値**に対して，



右のような図のことをいいます．したがって，まず，この5つの値を定義に従って求められることが必要です(⇨132)．

箱ひげ図はヒストグラム(⇨130)ほどデータの様子を詳しく表しているわけではありませんが，度数分布表をつくる必要もないので，そのおおまかな様子は簡単に知ることができます．

(2) 散らばり度合いは**四分位範囲** Q_3-Q_1 か**四分位偏差** $\dfrac{Q_3-Q_1}{2}$ の大小で比べるので，箱ひげ図の**長方形の横の辺の長さ**でわかります．

解答

(1) **(K線について)**

データを小さい順に並べかえると，

0.8，1.1，1.5，1.9，2.1，2.1，2.2，2.7，2.8，3.0，3.5，4.0，5.1，5.1，5.2

↓ m　↓ Q_1　↓ Q_2　↓ Q_3　↓ M

よって，$m=0.8$，$M=5.2$，$Q_1=1.9$，$Q_2=2.7$，$Q_3=4.0$

（I 線について）

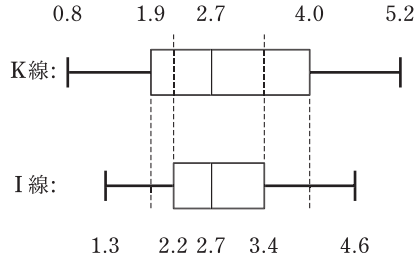
データを小さい順に並べかえると，

1.3, 1.4, 1.8, 2.2, 2.3, 2.5, 2.6, 2.7, 2.8, 2.9, 3.1, 3.4, 4.2, 4.3, 4.6

↓ m　↓ Q_1　↓ Q_2　↓ Q_3　↓ M

よって，$m=1.3$，$M=4.6$，$Q_1=2.2$，$Q_2=2.7$，$Q_3=3.4$

これより，K 線と I 線の箱ひげ図は，図のようになる．



(2)　K線について，$Q_3-Q_1=2.1$　　I 線について，$Q_3-Q_1=1.2$

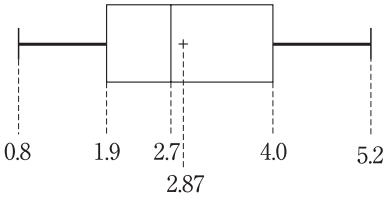
よって，K線の方が駅間距離の散らばり度合いが大きいといえる．

注　箱ひげ図に，平均値をかき込むことがあります．

このときは，記号「＋」を使います．

たとえば，K 線の平均値は小数第 3 位を四捨五入すると 2.87 になります．

だから，以下のような箱ひげ図になります．



ポイント

箱ひげ図は，データの次の 5 つの値を求める

①最大値　②最小値　③第 1 四分位数

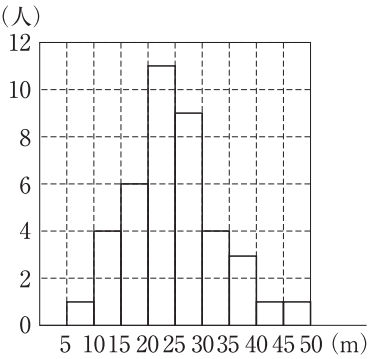
④第 2 四分位数　⑤第 3 四分位数

演習問題 134

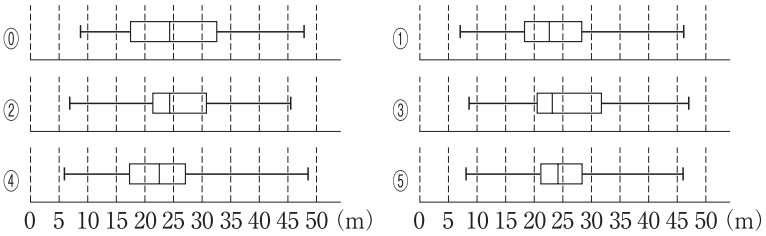
132 のデータを使ってA君，B君それぞれについて箱ひげ図をかけ．

135 ヒストグラムと箱ひげ図

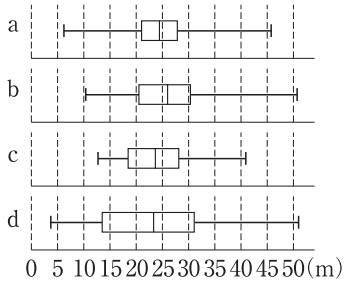
ある高校3年生1クラスの生徒40人について，ハンドボール投げの飛距離のデータを取った．右の図は，このクラスで最初に取ったデータのヒストグラムである．



(1) このデータを箱ひげ図にまとめたとき，右図のヒストグラムと矛盾するものはどれか．理由を述べて，すべて求めよ．



(2) 後日，このクラスでハンドボール投げの記録を取り直した．次に示したA～Dは，最初に取った記録から今回の記録への変化の分析結果を記述したものである．a～dの各々が今回取り直したデータの箱ひげ図となる場合に，⓪～③の組合せのうち分析結果と箱ひげ図が矛盾するものはどれか．理由を述べて，すべて求めよ．



⓪ A—a　① B—b
② C—c　③ D—d

A：どの生徒の記録も下がった．
B：どの生徒の記録も伸びた．
C：最初に取ったデータで上位 $\frac{1}{3}$ に入るすべての生徒の記録が伸びた．

D：最初に取ったデータで上位 $\frac{1}{3}$ に入るすべての生徒の記録は伸び，下位 $\frac{1}{3}$ に入るすべての生徒の記録は下がった．

精講

(1)　箱ひげ図に現れる代表値は，134にあるように，最小値 m，第1四分位数 Q_1，第2四分位数 Q_2，第3四分位数 Q_3，最大値 M の5つですが，ヒストグラムでは個々のデータがわからないので，この5つの値を正確に知ることはできません．しかし，**ある程度の幅をもって**知ることはできます(⇨131)．たとえば，「m は5点から10点の間」というように．

よって，ヒストグラムから m，Q_1，Q_2，Q_3，M の属する階級を読みとり，箱ひげ図と比べていくことになりますが，このような選択式の問題では，ヒストグラムと⓪，ヒストグラムと①，…と比べていくのではなく，まず，m について，⓪～⑤を比べて，不適切なものを答から外し，以下，M について，Q_1 について，…と考えていく方が**時間をムダにしないで答を選べる**ことも知っておきましょう．

(2)　⓪～③まで，「すべての生徒」に対する記述になっています．(1)でも述べたように，箱ひげ図では個々のデータが正確にはわからないので，分析と箱ひげ図が矛盾していない可能性があるが，断定できない場合があります．

ここで注意したいのは，矛盾していない(≒正しい)と断定できなくても，必ずしも矛盾しているわけではないことです．

解　答

(1)　**(m について)**

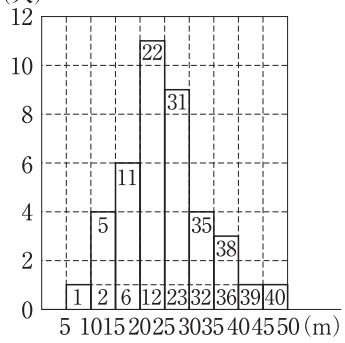
m は 5 m ～ 10 m の階級にあるので，すべて適する．

(M について)

M は 45 m ～ 50 m の階級にあるので，すべて適する．

(Q_1 について)

Q_1 は10人目と11人目が属する階級．



すなわち，15 m ～ 20 m の階級にある．

よって，②，③，⑤がヒストグラムと矛盾する．

（Q_2 について）

Q_2 は，20 人目と 21 人目が属する階級．

すなわち，20 m ～ 25 m の階級にある．よって，すべて適する．

（Q_3 について）

Q_3 は，30 人目と 31 人目が属する階級．

すなわち，25 m ～ 30 m の階級にある．

よって，⓪，②，③がヒストグラムと矛盾する．

以上のことより，⓪，②，③，⑤がヒストグラムと矛盾する．

注 センター試験のようなマーク式では，もう少し時間が節約できる．最初に，5 つの代表値の各々について，すべての箱ひげ図で同じ階級に存在するものは調べる対象からはずしてよい．

だから，本問の場合，m，M，Q_2 についてはチェック不要で，消費時間を $\frac{2}{5}$ に節約できる．

(2) **（A－a について）**

前のデータでは，第 1 四分位数は 15 m ～ 20 m の階級にあるが，新しいデータでは，第 1 四分位数が 20 m ～ 25 m の階級にある．

よって，下位 $\frac{1}{4}$ の生徒の中に記録が伸びた生徒がいる．

∴ 矛盾する

（B－b について）

前と後のデータでは，最小値，第 1 四分位数，第 2 四分位数，第 3 四分位数，最大値のすべてが属する階級が上がっているが，これだけでは，すべての生徒の記録が伸びたかどうか判断できないので，矛盾しているとはいえない．

（C－c について）

前と後のデータでは，最大値の属する階級が下がっているので，上位 $\frac{1}{3}$ に入る生徒の少なくとも 1 人は記録が下がっている．

∴ 矛盾する

（D－d について）

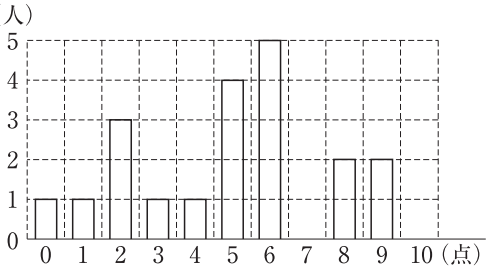
前と後のデータでは，最小値と第 1 四分位数の属する階級は下がり，

最大値と第3四分位数の属する階級は上がっているが，これだけでは，上位 $\frac{1}{3}$ に入るすべての生徒の記録が伸び，下位 $\frac{1}{3}$ の生徒の記録が下がったかどうか判断できないので，矛盾しているとはいえない．

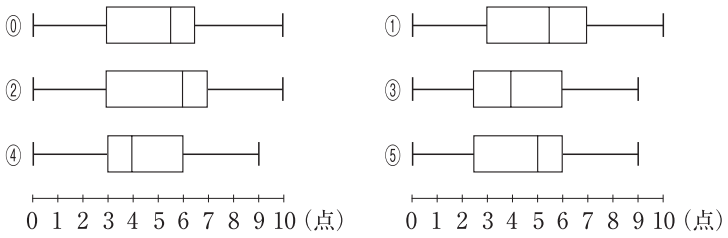
よって，矛盾するのは⓪と②である．

演習問題 135

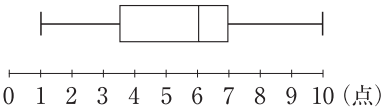
20人の生徒が10点満点のテストを受けた．そのデータを棒グラフで表すと右図のようになった．



(1) このテストの得点の箱ひげ図は下のどれか．理由を述べて答えよ．



(2) 後日，このテストのデータが間違っていることがわかり，再集計し，箱ひげ図を作り直したら，右図のようになった．



修正前と修正後の箱ひげ図を比較して，分析結果としてつねに正しいものは次のどれか．理由をつけて答えよ．

⓪ 得点の修正後の平均値は修正前の平均値より上がった．

① 得点の修正前と比較すると，少なくとも2人の得点が変化した．

② 得点の修正後のデータのばらつきは修正前に比べて大きくなった．

③ ⓪～②の中につねに正しいといえるものはない．

136 分散・標準偏差

次のデータはA君，B君2人の10回分のテストの結果である．

回	1	2	3	4	5	6	7	8	9	10
A君（点）	1	3	2	1	6	9	2	1	7	8
B君（点）	6	7	8	10	6	9	8	7	9	10

(1)　A君，B君それぞれの平均値，分散，標準偏差を求めよ．

(2)　(1)の結果から得点がより安定しているのはどちらといえるか．

精講

(1)　132でデータの散らばり度合いを判断する指標として四分位偏差を学びましたが，より正確な散らばり度合いを示す指標として，**分散**と**標準偏差**という数値を考えます．

（**分散**）　n 個のデータ x_1，x_2，…，x_n について，その平均値を $\overline{x}$ とするとき，

$\dfrac{1}{n}\{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2\}$ で表される値を**分散**といい，s^2 で表す．

（**標準偏差**）　分散 s^2 の正の平方根 s を**標準偏差**という．

注　分散も標準偏差もデータの散らばり度合いを表していますが，分散はデータを2乗するので単位が変わり，演算に不都合が生じます．このため標準偏差を考えるのです．

(2)　得点が安定しているとは，散らばり度合いが小さい，すなわち，**分散（標準偏差でもよい）が小さい**ことを指します．

解　答

(1)　A君の平均値，分散，標準偏差をそれぞれ，$\overline{x}_a$，s_a^2，s_a

B君も同様に，$\overline{x}_b$，s_b^2，s_b とおく．

$$\overline{x}_a=\frac{1}{10}(1+3+2+1+6+9+2+1+7+8)=\frac{40}{10}=\mathbf{4}\ (\text{点})$$

$$s_a^2=\frac{1}{10}\{(4-1)^2+(4-3)^2+(4-2)^2+(4-1)^2+(4-6)^2+(4-9)^2$$
$$+(4-2)^2+(4-1)^2+(4-7)^2+(4-8)^2\}$$

$$=\frac{1}{10}(9+1+4+9+4+25+4+9+9+16)=\frac{90}{10}=\mathbf{9}$$

$\therefore\quad s_a=\mathbf{3}$ (点)

$$\overline{x}_b=\frac{1}{10}(6+7+8+10+6+9+8+7+9+10)=\frac{80}{10}=\mathbf{8}\ (\text{点})$$

$$s_b{}^2=\frac{1}{10}\{(8-6)^2+(8-7)^2+(8-8)^2+(8-10)^2+(8-6)^2+(8-9)^2$$
$$+(8-8)^2+(8-7)^2+(8-9)^2+(8-10)^2\}$$
$$=\frac{1}{10}(4+1+4+4+1+1+1+4)=\frac{20}{10}=\mathbf{2}$$

$\therefore\quad s_b=\sqrt{\mathbf{2}}$ (点)

(2) $s_a>s_b$ だから，B君の方が安定している．

階級値	度　数
x_1	f_1
x_2	f_2
⋮	⋮
x_n	f_n
計	n

注　度数分布表から，標準偏差 s を求めるときは階級値 (⇨130) をデータと考えて，次の式で求めます．

(⇨**演習問題 136**)

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

ポイント

n 個のデータ x_1，x_2，…，x_n に対して，標準偏差 s は，

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

で表される

注　**偏差値**については142を参照してください．

演習問題 136

右表は，A，B 2クラスの身長についての度数分布表である．

それぞれのクラスについて平均値，分散，標準偏差を求め，身長の散らばり度合いはどちらが大きいか答えよ．

身長 (cm)	A	B
145以上155未満	5	1
155～165	6	4
165～175	4	12
175～185	4	2
185～195	1	1
計	20	20

137 計算の工夫

次のデータは5人のハンドボール投げの記録である．

28，a，24，b，c （単位はm）

このデータでは，次の4つの性質が成りたっている。

(ア) $24<a<28<b<c$

(イ) 第3四分位数は33 m

(ウ) 平均値は29 m

(エ) 分散は14

このとき，a，b，c の値を求めよ．

精講

文字が3つありますので，第3四分位数，平均値，分散の定義に従って等式を3つつくり，連立方程式を解けばよいだけですが，数値が大きいので，計算まちがいが心配です．

そこで，平均値がわかっているので，すべてのデータから29 mを引いた新しいデータを考えることで，**計算量を減らす工夫**を学びます．

解　答

与えられたデータから29 mを引いた数を新しいデータとして考える．

すなわち，小さい順に，

-5，$a-29$，-1，$b-29$，$c-29$

を考える．

$a'=a-29$，$b'=b-29$，$c'=c-29$ とおく．

(イ)より，$\dfrac{b+c}{2}=33$ だから，$b+c=66$

$\therefore\quad b'+c'=8$ ……①

(ウ)より，$24+a+28+b+c=29\cdot5$

$\therefore\quad a+b+c=29\cdot5-52$

よって，$a'+b'+c'+29\cdot3=29\cdot5-52$

$\therefore\quad a'+b'+c'=29\cdot2-52$

$\therefore\quad a'+b'+c'=6$ ……②

㋓より，$(24-29)^2+(a-29)^2+(28-29)^2+(b-29)^2+(c-29)^2=14\cdot5$

$\therefore\quad a'^2+b'^2+c'^2=44$ ……③

①，②より，$a'=-2$，$c'=8-b'$

③に代入して，$4+b'^2+(8-b')^2=44$

$\therefore\quad 2b'^2-16b'+64-40=0$

$b'^2-8b'+12=0$

$(b'-2)(b'-6)=0$

$\therefore\quad b'=2$ または 6

$b'=2$ のとき，$c'=6$

$b'=6$ のとき，$c'=2$ であるが，

$b<c$ より，$b'<c'$ だから不適．

よって，$b'=2$，$c'=6$

以上のことより，$a=\mathbf{27}$，$b=\mathbf{31}$，$c=\mathbf{35}$

注　もし，元のデータのまま解答をつくると，でき上がる連立方程式は $b+c=66$，$a+b+c=93$，$(a-29)^2+(b-29)^2+(c-29)^2=44$ となります．定数項を比べてみると一目瞭然ですね．

参考　視力検査の数値のように，小数点以下を含むデータのときの工夫の仕方は，**141** で学びます．

演習問題 137

次のデータは 5 人の体重測定の結果である．

57，64，a，b，c　（単位は kg）

このデータに対して，次の 4 つの性質が成りたっている．

㋐　$57<a<b<64<c$

㋑　データの範囲は 10 kg

㋒　データの平均値は 62 kg

㋓　データの分散は 11.6

このとき，a，b，c の値を求めよ．

138 もう1つの分散の求め方

(1)　n 個のデータを x_1，x_2，…，x_n とし，このデータの平均値を $\overline{x}$，分散を ${s_x}^2$ で表すとき，分散

${s_x}^2=\dfrac{1}{n}\{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2\}$ は，

${s_x}^2=\dfrac{1}{n}({x_1}^2+{x_2}^2+\cdots+{x_n}^2)-(\overline{x})^2$ と表せることを示せ．

(2)　6個のデータ，x_1，x_2，x_3，x_4，x_5，x_6 がある．このデータの平均値を $\overline{x}$，分散を ${s_x}^2$ とするとき，$\overline{x}=2$，${s_x}^2=5$ であった．

このとき，新しいデータ，${x_1}^2$，${x_2}^2$，${x_3}^2$，${x_4}^2$，${x_5}^2$，${x_6}^2$ の平均値を求めよ．

精講

(1)　$(a-b)^2=a^2-2ab+b^2$ を考えると，

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

の登場が想像できます．

ポイントは **$-2x_1\overline{x}-2x_2\overline{x}-\cdots-2x_n\overline{x}$ の処理**にあります．

(2)　ほしいものは，$\dfrac{{x_1}^2+{x_2}^2+{x_3}^2+{x_4}^2+{x_5}^2+{x_6}^2}{6}$，

すなわち，${x_1}^2+{x_2}^2+{x_3}^2+{x_4}^2+{x_5}^2+{x_6}^2$．

わかっているものは，$\overline{x}\left(=\dfrac{x_1+x_2+x_3+x_4+x_5+x_6}{6}\right)$ と ${s_x}^2$ ですから，

$\overline{x}$ と ${s_x}^2$ と ${x_1}^2+{x_2}^2+{x_3}^2+{x_4}^2+{x_5}^2+{x_6}^2$ をつなぐ

ことを考えます．

解　答

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

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

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

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

$\therefore \quad s_x^2=\dfrac{1}{n}(x_1^2+x_2^2+\cdots+x_n^2)-(\overline{x})^2$

(2) $s_x^2=\dfrac{1}{6}(x_1^2+x_2^2+x_3^2+x_4^2+x_5^2+x_6^2)-(\overline{x})^2$ だから

$$\frac{x_1^2+x_2^2+x_3^2+x_4^2+x_5^2+x_6^2}{6}=s_x^2+(\overline{x})^2$$
$$=5+2^2=9$$

よって，x_1^2，x_2^2，x_3^2，x_4^2，x_5^2，x_6^2 の平均値は **9**

注 2つの分散の公式はどんな違いがあるのでしょうか？

扱うデータが具体的な数値の場合，各データ x_1，x_2，…，x_n が正の値であることが普通ですから

$(x_1-\overline{x})^2$ を x_1^2 と比べると，$(x_1-\overline{x})^2<x_1^2$

が成りたち，前者の公式の方が負担が軽くなります．

ところが，各データ x_1，x_2，…，x_n が整数であっても，$\overline{x}$ は小数になるのが普通です．そうすると，

$x_1-\overline{x}$，$x_2-\overline{x}$，…，$x_n-\overline{x}$ は小数で，

前者は小数の平方を n 回することになり，

後者は $(\overline{x})^2$ の部分1回だけで済みます．

どちらも大切で，使い分けできることが必要です．

ポイント

n 個のデータ x_1，x_2，…，x_n の分散 s_x^2 を求める公式は，$\overline{x}$ を平均値として

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

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

の2つがある

演習問題 138

8個の正方形 C_1，C_2，…，C_8 があり，その1辺の長さの平均は3で分散は4である．このとき，8個の正方形の面積の平均を求めよ．

139 代表値の変化（データの合算）

2つのグループA，Bに対して，10点満点のテストを実施した．Aグループは5人で，Bグループは10人である．

Aグループの平均を $\overline{a}$，分散を s_a^2，Bグループの平均を $\overline{b}$，分散を s_b^2 とするとき，$\overline{a}=8.2$，$s_a^2=5.2$，$\overline{b}=7.9$，$s_b^2=4.5$ であった．この15人の成績を合わせたときの平均を $\overline{x}$，分散を s_x^2 とする．ただし，これらの値はすべて正確な値であり，四捨五入されていないものとする．

(1) Aグループの得点を a_1，a_2，…，a_5，Bグループの得点を b_1，b_2，…，b_{10} とするとき，$a_1+a_2+\cdots+a_5$，$b_1+b_2+\cdots+b_{10}$ の値を求め，$\overline{x}$ を求めよ．

(2) $a_1^2+a_2^2+\cdots+a_5^2$，$b_1^2+b_2^2+\cdots+b_{10}^2$ の値を求め，s_x^2 を求めよ．ただし，小数第2位を四捨五入せよ．

精講

(1) $\overline{x}=\dfrac{a_1+a_2+\cdots+a_5+b_1+b_2+\cdots+b_{10}}{15}$ と表されますので

$a_1+a_2+\cdots+a_5$ と $b_1+b_2+\cdots+b_{10}$ の値が必要になります．

(2) 分散の定義によれば

$$s_x^2=\frac{(a_1-\overline{x})^2+(a_2-\overline{x})^2+\cdots+(a_5-\overline{x})^2+(b_1-\overline{x})^2+(b_2-\overline{x})^2+\cdots+(b_{10}-\overline{x})^2}{15}$$

と表されますが，誘導されているのは，

$a_1^2+a_2^2+\cdots+a_5^2$ と $b_1^2+b_2^2+\cdots+b_{10}^2$ の値

で，これらは，s_x^2 の右辺を展開すると確かにその一部として登場します．

しかし，まともに展開すると，45もの項が出てくるので，何か上手に手段を考えたい．そのためには，分散のもう1つの求め方（⇨138）を知っておく必要があります．

すなわち，言葉でいうと，**分散＝(2乗の平均)−(平均)^2** で，

式で表すと，

$$\boldsymbol{s_x^2=\frac{1}{15}(a_1^2+a_2^2+\cdots+a_5^2+b_1^2+b_2^2+\cdots+b_{10}^2)-(\overline{x})^2}$$

です．

解　答

(1)　$a_1+a_2+\cdots+a_5=\overline{a}\times5$

$\therefore\quad a_1+a_2+\cdots+a_5=8.2\times5=\mathbf{41}$

$b_1+b_2+\cdots+b_{10}=\overline{b}\times10$

$\therefore\quad b_1+b_2+\cdots+b_{10}=7.9\times10=\mathbf{79}$

よって，

$$\overline{x}=\frac{(a_1+a_2+\cdots+a_5)+(b_1+b_2+\cdots+b_{10})}{15}=\frac{41+79}{15}=\frac{120}{15}=8$$

$\therefore\quad \overline{x}=\mathbf{8}$

(2)　$s_a{}^2=\dfrac{1}{5}(a_1{}^2+a_2{}^2+\cdots+a_5{}^2)-(\overline{a})^2$ だから

$a_1{}^2+a_2{}^2+\cdots+a_5{}^2=5\{s_a{}^2+(\overline{a})^2\}=5(5.2+67.24)=\mathbf{362.2}$

$b_1{}^2+b_2{}^2+\cdots+b_{10}{}^2=10\{s_b{}^2+(\overline{b})^2\}=10(4.5+62.41)=\mathbf{669.1}$

よって，$s_x{}^2=\dfrac{1}{15}(a_1{}^2+a_2{}^2+\cdots+a_5{}^2+b_1{}^2+b_2{}^2+\cdots+b_{10}{}^2)-(\overline{x})^2$

$$=\frac{1}{15}(362.2+669.1)-64=\frac{1031.3-960}{15}=4.75\cdots$$

小数第 2 位を四捨五入して，$s_x{}^2=\mathbf{4.8}$

ポイント

n 個のデータ x_1，x_2，$\cdots$，x_n の平均を $\overline{x}$，分散を $s_x{}^2$ とするとき，

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

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

演習問題 139

4 人のグループＡと 6 人のグループＢがあって，合計 10 人がテストを受けた．

Ａグループの平均を $\overline{a}$，分散を $s_a{}^2$，Ｂグループの平均を $\overline{b}$，分散を $s_b{}^2$ とするとき，$\overline{a}=8.0$，$s_a{}^2=4.0$，$\overline{b}=7.0$，$s_b{}^2=5.0$ であった．

このとき，10 人全体の平均 $\overline{x}$ と分散 $s_x{}^2$ を求めよ．

140 代表値の変化（データの追加）

10人の生徒が10点満点のテストを受けた．
得点の低い順に並べたデータを x_1，x_2，…，x_{10} とする．
最低点の生徒は合格点に達しなかったので，翌日追試を受けて合格点をとった．追試前の平均，分散をそれぞれ $\overline{x}$，s_x^2，追試後の平均，分散をそれぞれ，$\overline{y}$，s_y^2 とするとき，次の問いに答えよ．

(1)　$\overline{x}$ と $\overline{y}$ の大小を判断せよ．

(2)　$\overline{x}=7$，$s_x^2=3.4$ とする．
追試を受けた生徒の得点が3点から5点になったとき $\overline{y}$ と s_y^2 の値を求めよ．

精講

データに変更があると，代表値（平均，分散，四分位数など）も変化するのが普通ですが，変化の様子を(1)のように，大きくなる，小さくなる，という観点で判断する場合と，(2)のように，代表値の変化で判断する場合の2つがあります．どちらも大切な判断法です．

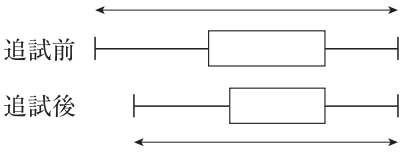
(1)では，箱ひげ図や，定義の式のイメージが有効で，

(2)では，定義に従ってキチンと計算することが必要です．

解　答

(1)　最低点だった生徒の得点が増えているので，10人分の得点の総和は増える．
よって，平均点は追試後の方が高くなる．
$\therefore\quad \overline{x}<\overline{y}$

◀定義の式で分母が不変だから分子の増減を考えている．



注　各四分位数の変化や，分散の変化は，これだけの情報では判断できません．

(2)　追試を受けた生徒の得点が x_1' のとき，$x_1'=x_1+2$

$$\therefore\quad \overline{y}=\frac{x_1'+x_2+\cdots+x_{10}}{10}=\frac{x_1+x_2+\cdots+x_{10}+2}{10}=\overline{x}+0.2=\mathbf{7.2}$$

$$s_y{}^2=\frac{1}{10}(x_1{}'^2+x_2{}^2+\cdots+x_{10}{}^2)-(\overline{y})^2$$ ◀138

$$=\frac{1}{10}\{(x_1+2)^2+x_2{}^2+\cdots+x_{10}{}^2\}-(\overline{y})^2$$

$$=\frac{1}{10}(x_1{}^2+x_2{}^2+\cdots+x_{10}{}^2+4x_1+4)-(\overline{y})^2$$

$$=\frac{1}{10}(x_1{}^2+x_2{}^2+\cdots+x_{10}{}^2)-(\overline{x})^2+(\overline{x})^2-(\overline{y})^2+\frac{2(x_1+1)}{5}$$

$$=s_x{}^2+(\overline{x}+\overline{y})(\overline{x}-\overline{y})+\frac{2}{5}(3+1)$$

$$=s_x{}^2-14.2\times0.2+1.6$$

$$=s_x{}^2-2.84+1.6=3.4-1.24=\mathbf{2.16}$$

ポイント

データが変化したときの代表値の変化は，

・性質から判断する
・代表値を求めて判断する

の 2 つの場合があり，前者は箱ひげ図や定義の式のイメージから判断する

演習問題 140

9 人の生徒が 10 点満点のテストを受けた．
このテストの得点を x_1，x_2，…，x_9 とする．
翌日，1 人欠席の生徒がテストを受け，得点は 9 点であった．
最初の 9 人分の平均，分散をそれぞれ $\overline{x}$，$s_x{}^2$ とすると
$\overline{x}=6$，$s_x{}^2=4$ であった．10 人分の平均 $\overline{y}$ と分散 $s_y{}^2$ を求めよ．

141 代表値の変化（変量変換）

(1)　平均が $\overline{x}$，分散が s_x^2 である n 個のデータ x_1，x_2，…，x_n と平均が $\overline{y}$，分散が s_y^2 である n 個のデータ y_1，y_2，…，y_n があり，2つの変量の間には，a，b を定数として $y_i=ax_i+b$ $(i=1, 2, 3, \cdots, n)$ の関係があるとする．

このとき，次の問いに答えよ．

(ア)　$\overline{y}=a\overline{x}+b$ が成りたつことを示せ．

(イ)　$s_y^2=a^2s_x^2$ が成りたつことを示せ．

(2)　次のデータは5人の通学距離の測定結果である．

2.6，1.4，1.8，0.7，3.0　（単位は km）

このデータの平均 $\overline{x}$ と分散 s_x^2 を $y=10x-20$ を利用して求めよ．

精講　この考え方は，137で話した内容を一般化したものです．厳密には数学Bの範囲ですが，これを知っておくと，大きなデータ，小さなデータを扱うときの計算ミスの確率が下がります．センター試験のような**答だけでよい問題**では，特に有効です．

解　答

(1)　(ア)　$\overline{y}=\dfrac{1}{n}(y_1+y_2+\cdots+y_n)$

$=\dfrac{1}{n}\{(ax_1+b)+(ax_2+b)+\cdots+(ax_n+b)\}$

$=\dfrac{1}{n}\{a(x_1+x_2+\cdots+x_n)+nb\}$

$=\dfrac{1}{n}(a\cdot n\overline{x}+nb)$　　◀ $\overline{x}=\dfrac{x_1+x_2+\cdots+x_n}{n}$

$=a\overline{x}+b$

(イ)　$s_y^2=\dfrac{1}{n}(y_1^2+y_2^2+\cdots+y_n^2)-(\overline{y})^2$　◀ 138

$=\dfrac{1}{n}\{(ax_1+b)^2+(ax_2+b)^2+\cdots+(ax_n+b)^2\}-(a\overline{x}+b)^2$

$$=\frac{1}{n}\{a^2(x_1^2+x_2^2+\cdots+x_n^2)+2ab(x_1+x_2+\cdots+x_n)+nb^2\}$$
$$-\{a^2(\overline{x})^2+2ab\overline{x}+b^2\}$$
$$=a^2\cdot\frac{1}{n}(x_1^2+x_2^2+\cdots+x_n^2)+\frac{1}{n}\cdot 2ab\cdot n\overline{x}+b^2-a^2(\overline{x})^2$$
$$-2ab\overline{x}-b^2$$
$$=a^2\cdot\frac{1}{n}(x_1^2+x_2^2+\cdots+x_n^2)+2ab\overline{x}+b^2-a^2(\overline{x})^2-2ab\overline{x}-b^2$$
$$=a^2\left\{\frac{1}{n}(x_1^2+x_2^2+\cdots+x_n^2)-(\overline{x})^2\right\}=a^2s_x^2$$

よって，$s_y^2=a^2s_x^2$

(2) 5つのデータを順に x_1，x_2，x_3，x_4，x_5 とし，

$y_i=10x_i-20$ $(i=1,\ 2,\ 3,\ 4,\ 5)$ で変換すると

$y_1=6$，$y_2=-6$，$y_3=-2$，$y_4=-13$，$y_5=10$

よって，$\overline{y}=\dfrac{6+(-6)+(-2)+(-13)+10}{5}=-1$

◀この計算がラクになる

$\therefore\quad -1=10\overline{x}-20$ より，$\overline{x}=\mathbf{1.9}$ **(km)**

また，$s_y^2=\dfrac{1}{5}\{6^2+(-6)^2+(-2)^2+(-13)^2+10^2\}-(\overline{y})^2$

$=\dfrac{1}{5}(36+36+4+169+100)-(-1)^2=68$ だから

$68=10^2s_x^2\qquad\therefore\quad s_x^2=\mathbf{0.68}$

ポイント

平均が $\overline{x}$，分散 s_x^2 のデータを $y=ax+b$ で変換すると，y の平均 $\overline{y}$，分散 s_y^2 はそれぞれ

$\overline{y}=a\overline{x}+b$，$s_y^2=a^2s_x^2$

で表される

演習問題 141

次のデータは5人の身長の測定結果である．

166，158，177，187，162 （単位は cm）

このデータの平均 $\overline{x}$ と分散 s_x^2 を $y=x-167$ を利用して変量を変換して求めよ．

142 偏差値

ある会社の入社試験で，国語と数学の試験が行われた．

国語の平均を $\overline{x}$，標準偏差を s_x，数学の平均を $\overline{y}$，標準偏差を s_y とするとき，$\overline{x}=62$，$s_x=15$，$\overline{y}=55$，$s_y=20$ であった．

(1)　受験者Aは，国語，数学ともに80点をとった．それぞれの科目の偏差値を求めよ．

ただし，平均が m，標準偏差が σ のデータに対して，変量 x の偏差値は $\dfrac{x-m}{\sigma}\times 10+50$ で求められる値である．

(2)　2人の受験者A，Bに対して，得点は右表のようになった．科目間の難易度を反映させるために，得点の合計ではなく，偏差値の合計で合否を決めることになった．

合格しやすいのはA，Bのどちらか．

	A	B
国語	80	74
数学	80	87
合計	160	161

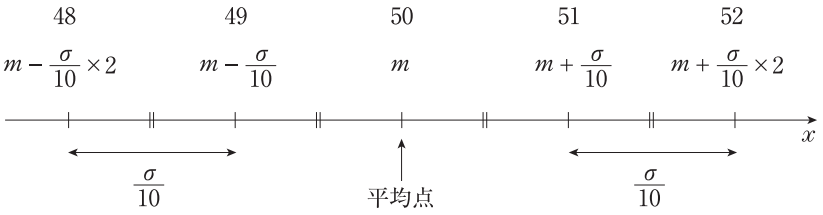
精講

受験生には，切っても切れない数値である偏差値がテーマです．

受験生でない人でも，この単語を聞いたことがないという人はいないと思いますが，どうやって求めているのか，どんな意味をもっているのかを知らないで，「偏差値が65だから…」などという会話を耳にします．

また，世間では，偏差値は悪者のようにいわれているという側面も否定できません．入試ではこの問題のように定義の式が与えられるので，覚えておく必要はありませんが，せめて異質な2つの数値に対する評価方法の1つであることは知っておいてほしいものです．

定義の式から得られる偏差値のイメージは下図のようなものです．



解　答

(1)　国語の偏差値は

$$\frac{80-62}{15}\times 10+50=\frac{18}{15}\times 10+50=\mathbf{62}$$

数学の偏差値は

$$\frac{80-55}{20}\times 10+50=\frac{25}{20}\times 10+50=\mathbf{62.5}$$

(2)　(1)より，A の偏差値の合計は $62+62.5=124.5$

次に，B の国語の偏差値は

$$\frac{74-62}{15}\times 10+50=58$$

B の数学の偏差値は

$$\frac{87-55}{20}\times 10+50=66$$

よって，B の偏差値の合計は $58+66=124$

以上のことより，A の方がより合格に近い．

参考　(2)では，得点の合計ではBの方が勝っているのに，偏差値では，A の方が勝っています．これは，**標準偏差の小さい方が高偏差値になりやすい**からです．精講の図によると，数直線上で，$\frac{\sigma}{10}$ が小さい方が，偏差値を 1 上げるのに必要な得点が少なくてすむということです．

演習問題 142

2 科目入試の大学を A，B の 2 人が受験した．科目 X，科目 Y の得点は右表のようであった．

	A	B
X	96	88
Y	90	99
合計	186	187

X の平均を $\overline{x}$，標準偏差を s_x，
Y の平均を $\overline{y}$，標準偏差を s_y とするとき，
$\overline{x}=72$，$s_x=16$，$\overline{y}=84$，$s_y=24$ であった．

2 科目の偏差値の合計で順位が決まるとき，A，B のどちらが上位の成績といえるか．

143 散布図と相関

次の表は12人の生徒に行った10点満点で2回ずつ実施したA，B2科目のテストの結果である．

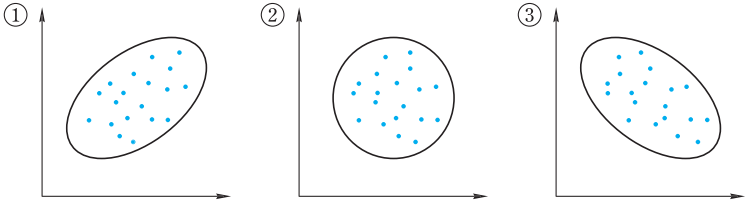
番号		1	2	3	4	5	6	7	8	9	10	11	12
1回目	A	1	9	9	2	7	4	6	2	8	8	6	4
	B	4	5	7	1	8	6	7	6	10	9	5	4
2回目	A	3	9	5	2	7	4	6	1	7	2	5	3
	B	3	8	3	2	7	5	5	3	8	4	7	5

(1) 1回目，2回目それぞれについて，AとBの散布図をかけ．

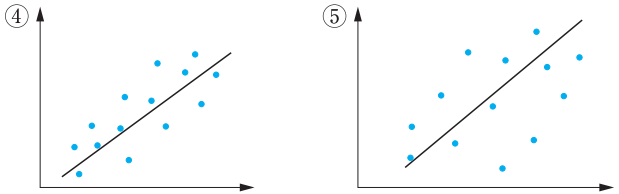
(2) (1)の散布図を利用して，1回目，2回目のどちらの相関が強いか判断せよ．

精講

(1) 2つのデータの間に関連性があるかどうかを調べるとき，散布図をかくとその雰囲気がつかめます．散布図のかき方は**座標の考え方と同じ**で，たとえば，1回目の1番の人の場合，座標平面上の点(1，4)に印をつけます．散布図が下図①のようなとき，**正の相関関係**がある，③のようなとき，**負の相関関係**がある，②のようなとき，**相関関係がない**とそれぞれいいます．

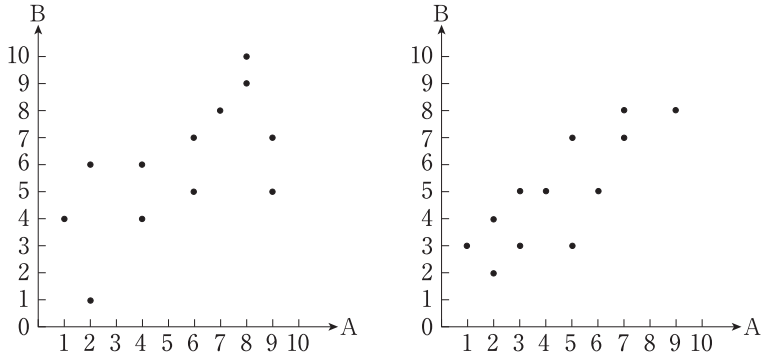


また，下図の④と⑤の散布図を比べると，④の方が，⑤より点が密集している感じがします．このようなとき，④の方が⑤より**相関が強い**といいます．



解　答

(1)　(1回目の散布図)　(2回目の散布図)



(2)　2回目の散布図の方が1回目の散布図に比べて点の密集感があるので，2回目のテストの方が相関が強いといえる．

参考　これはフンイキですから，密集度合を数値で表すとキチンと相関の強弱が数学らしく求められます．これについては145の相関係数で学びます．

ポイント　散布図を用いると，正確さはともかく，短時間で相関の強弱を知ることができる

演習問題 143

次の表は10人の生徒に行った10点満点で2回ずつ実施したA，B2科目のテストの結果である．

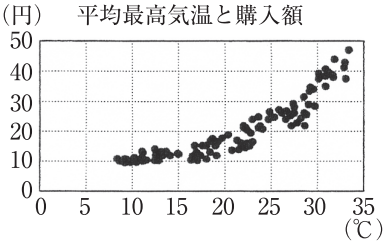
番号		1	2	3	4	5	6	7	8	9	10
1回目	A	5	6	2	6	1	4	2	4	3	2
	B	5	7	1	6	3	5	2	4	3	4
2回目	A	3	7	1	4	4	5	2	4	3	5
	B	5	6	2	6	3	8	3	2	1	4

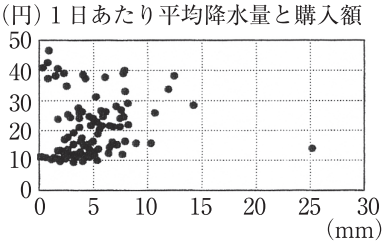
(1)　1回目，2回目それぞれについて，AとBの散布図をかけ．

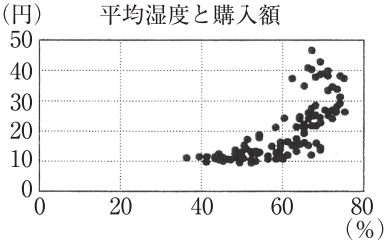
(2)　(1)の散布図を利用して，1回目，2回目のどちらの相関が強いか判断せよ．

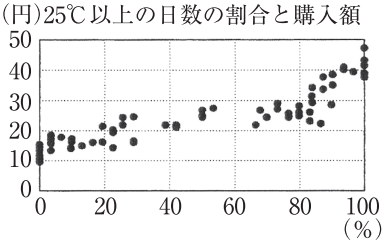
144 散布図（読みとり）

次の4つの散布図は，2003年から2012年までの120か月の東京の月別データをまとめたものである．それぞれ，1日の最高気温の月平均（以下，平均最高気温），1日あたり平均降水量，平均湿度，最高気温25℃以上の日数の割合を横軸にとり，各世帯の1日あたりアイスクリーム平均購入額（以下，購入額）を縦軸としてある．









次の⓪～④について，これらの散布図から正しいと読みとれるかどうか理由を付けて述べよ．

⓪ 平均最高気温が高くなるほど購入額は増加する傾向がある．

① 1日あたり平均降水量が多くなるほど購入額は増加する傾向がある．

② 平均湿度が高くなるほど購入額の散らばりは小さくなる傾向がある．

③ 25℃以上の日数の割合が80％未満の月は，購入額が30円を超えていない．

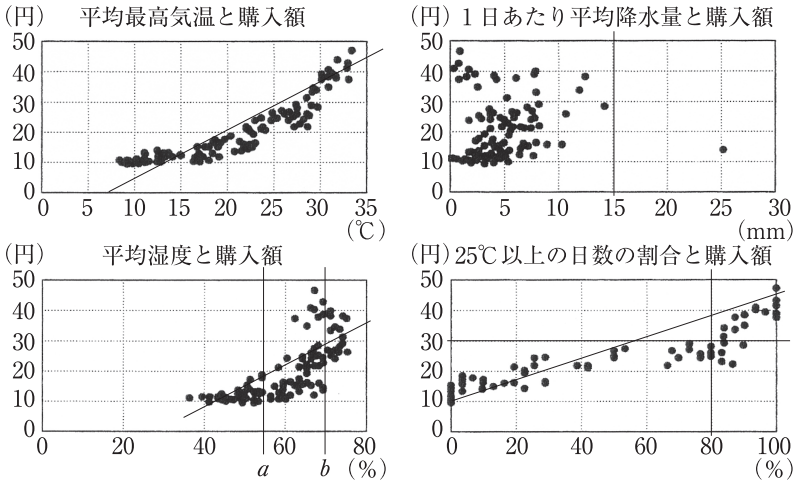
④ この中で正の相関があるのは，平均湿度と購入額の間のみである．

精講 散布図というのは，**2つのデータを座標のように点で表して，座標平面上にかき込んだもの**です（⇨143）．

だから，平均値や分散のようなデータの代表値を知ることはできません．しかし，様々な傾向を読みとることはできます．

実際の入試問題では，出題形式はこの問題の形になると思われます．カンで答えるのではなく，**根拠をもって**（＝理由をつけて）答えられるようになってください．

解　答



（⓪について）

左上図によると，点は右上がりの直線に沿って並んでいるので，正しいといえる．

（①について）

右上図によると，平均降水量が15 mmを超えても，アイスクリームはほとんど購入されていない．また，15 mmより小さいところでは，どの降水量に対しても，点は上から下までまんべんなく並んでいる．

よって，平均降水量が多くなったからといって，アイスクリームの平均購入額が増えるとはいえない．

よって，正しいとはいえない．

（②について）

左下図によると，2つの平均湿度a％とb％（$a<b$）のところで縦

線をひいてみると，a の線上よりも b の線上の方が点の存在する範囲が長い傾向がある．
したがって，平均湿度が高くなるとアイスクリームの平均購入額の散らばりは大きくなる．
よって，正しいとはいえない．

(③について)

右下図によると，80 % のところで縦線をひいて，その直線上にある一番上の点から横線をひく．縦線より左側の領域で，この横線より上側に点は存在しない．
よって，正しいといえる．

(④について)

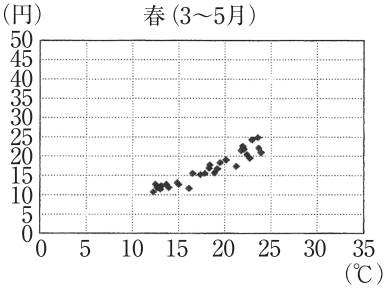
右上の散布図を除き，傾き正の直線上に沿って点が集まっている傾向があるので，正しいとはいえない．

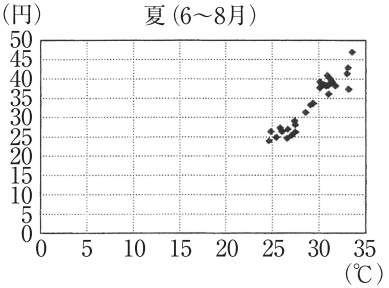
ポイント

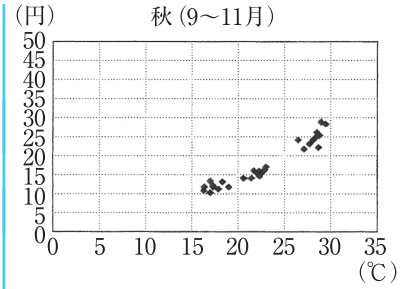
散布図から傾向を読みとる問題では，文章の表現に注意する
・～となる傾向がある　・～である　・～でない
・～のみ　・少なくとも～　・つねに～

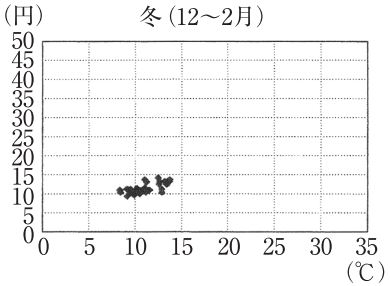
演習問題 144

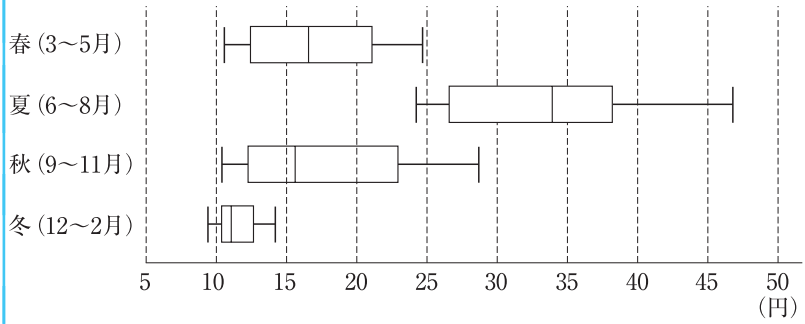
次の 4 つの散布図は，242 ページの散布図『平均最高気温と購入額』のデータを季節ごとにまとめたもので，その下にある 4 つの箱ひげ図は，購入額のデータを季節ごとにまとめたものである．











次の［ア］，［イ］に当てはまるものを，下の⓪〜⑧のうちから1つずつ選べ．ただし，解答の順序は問わない．

季節ごとの平均最高気温と購入額について，これらの図から読みとれることとして正しいものは，［ア］と［イ］である．

⓪　夏の購入額は，すべて25円を上回っている．

①　秋には平均最高気温が20℃以下で購入額が15円を上回っている月がある．

②　購入額の範囲が最も大きいのは秋である．

③　春よりも秋の方が，購入額の最大値は小さい．

④　春よりも秋の方が，購入額の第3四分位数は大きい．

⑤　春よりも秋の方が，購入額の中央値は大きい．

⑥　平均最高気温が25℃を上回っている月があるのは夏だけである．

⑦　購入額の四分位範囲が最も小さいのは春である．

⑧　購入額が35円を下回っている月は，すべて平均最高気温が30℃未満である．

145 共分散・相関係数

下の表は10人が参加した試合の1回戦と2回戦の各人の得点である．

番　号	1	2	3	4	5	6	7	8	9	10
1回戦 (x)	33	30	44	38	29	43	33	34	36	30
2回戦 (y)	37	34	44	35	30	41	33	38	41	37

(1)　1回戦，2回戦の平均値をそれぞれ $\overline{x}$，$\overline{y}$，分散を ${s_x}^2$，${s_y}^2$ とする．$\overline{x}$，$\overline{y}$，${s_x}^2$，${s_y}^2$ を求めよ．

(2)　共分散 s_{xy} を求め，相関係数 r を求めよ．ただし，小数第3位を四捨五入せよ．

精講

(1)　平均値と分散は **136** で学んだ定義通り計算します．

(2)　n 個のデータの組 $(x_1,\ y_1)$，$(x_2,\ y_2)$，…，$(x_n,\ y_n)$ に対して $(x_i-\overline{x})(y_i-\overline{y})$ の平均値，すなわち

$$\frac{1}{n}\{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})\}$$

を x と y の**共分散**といい，記号 s_{xy} で表します．

また，s_x，s_y，s_{xy} に対して $r=\dfrac{s_{xy}}{s_x s_y}$ を x と y の変量の**相関係数**といいます．

相関係数 r は $-1\leqq r\leqq 1$ が成りたち，r が1に近づくほど**強い正の相関がある**といい，-1 に近づくほど**強い負の相関がある**といいます．

143 で学んだ散布図では，2つのデータの相関を雰囲気で判断しましたが，これを数値化したものが相関係数です．

解　答

(1)　$\overline{x}=\dfrac{1}{10}(33+30+44+38+29+43+33+34+36+30)=\mathbf{35}$（点）

${s_x}^2=\dfrac{1}{10}\{(-2)^2+(-5)^2+9^2+3^2+(-6)^2+8^2+(-2)^2+(-1)^2+1^2+(-5)^2\}$

$=25$　　$\therefore$　${s_x}^2=\mathbf{25}$

$\overline{y}=\dfrac{1}{10}(37+34+44+35+30+41+33+38+41+37)=\mathbf{37}$（点）

$s_y^2=\frac{1}{10}\{0^2+(-3)^2+7^2+(-2)^2+(-7)^2+4^2+(-4)^2+1^2+4^2+0^2\}=16$

$\therefore\quad s_y^2=\mathbf{16}$

(2) $s_{xy}=\frac{1}{10}\{(-2)\cdot0+(-5)(-3)+9\cdot7+3\cdot(-2)+(-6)(-7)+8\cdot4$

$+(-2)(-4)+(-1)\cdot1+1\cdot4+(-5)\cdot0\}=\mathbf{15.7}$

よって，$r=\frac{s_{xy}}{s_x s_y}=\frac{15.7}{5\times4}=0.785$

小数第3位を四捨五入して，$r=\mathbf{0.79}$

注 1つ1つのデータが大きいので，$\overline{x}$，$\overline{y}$ を求めるとき計算まちがいが心配です．このようなとき，次のような操作をすると，少し計算の負担が軽くなります(この考え方を**仮平均**といいます)．

10個の y のデータをみると，35点以上のデータが7個，35点より小さいデータが3個あるので，35点が0点になるような新しいデータ y' を考えます(⇨137，141)．

y	37	34	44	35	30	41	33	38	41	37
y'	+2	−1	+9	0	−5	+6	−2	+3	+6	+2

y' の平均 $\overline{y'}$ は

$\overline{y'}=\frac{1}{10}(2-1+9-5+6-2+3+6+2)=\frac{9+3+6+2}{10}=2$

よって，y の平均は $35+2=37$(点)

ポイント

n 個のデータの組 $(x_1,\ y_1)$，$(x_2,\ y_2)$，…，$(x_n,\ y_n)$ について，x の平均を $\overline{x}$，y の平均を $\overline{y}$ とすると，

共分散 s_{xy} は

$s_{xy}=\frac{1}{n}\{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})\}$

で表され，x の分散を s_x^2，y の分散を s_y^2 で表すとき，相関係数 r は，$r=\frac{s_{xy}}{s_x s_y}$ で表される．このとき，

$-1\leqq r\leqq1$ が成りたつ

⇨演習問題 145 は 248 ページ

演習問題 145

次のデータは10人の右手(x)と左手(y)の各人の握力の測定結果である.

番　号	1	2	3	4	5	6	7	8	9	10
右手(x)	50	52	46	42	43	35	48	47	50	37
左手(y)	31	33	48	42	51	49	39	45	45	47

(kg)

(1)　x と y の平均 $\overline{x}$, $\overline{y}$ と分散 s_x^2, s_y^2 を求めよ.

(2)　共分散 s_{xy} を求め, 相関係数 r を求めよ. ただし, 小数第3位を四捨五入せよ.