1　データの分析

1.1　度数分布表とヒストグラム

身長や気温など人や物の特性を数量的に表すものを**変量**と言います。調査などで得られた変量の観測値や測定値の集まりを**データ**、データの個数のことを**データの大きさ**と言います。例えば、ある A 地点の平均気温を表すデータが単位を (℃) として

21.7, 19.9, 18.1, 15.8, 14.7, 13.3, 12.9, 13.7, 16.5, 18.8, 11.2, 10.5, 17.9, 14.1, 17.1, 14.9, 19.5, 15.5, 16.2, 15.2

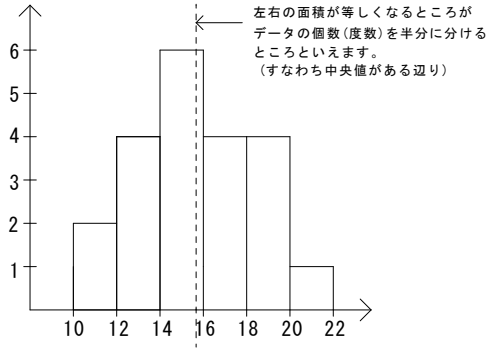
だった場合、データの大きさは 20 です。下の表にもこの 20 個のデータを用いました。

データの散らばりの様子を見る方法として**度数分布表**があります。度数分布表において、区切られた各区間を**階級**、各区間の幅を**階級の幅**、各階級に入るデータの値の個数を**度数**と言います。また、各階級の真ん中の値を**階級値**と言います。例えば 12 ℃以上 14 ℃未満の階級の階級値は 13 ℃です。また，各階級の度数の全体に対する割合を**相対度数**と言います。度数分布表を**ヒストグラム**を用いて視覚化することができます。ヒストグラムは横軸に階級、縦軸に度数をとって柱状に表したもので、度数が柱の面積と比例します。

度数分布表

階級 (℃)	階級値	度数	相対度数
10 以上 12 未満	11	2	0.1
12〜14	13	3	0.15
14〜16	15	6	0.3
16〜18	17	4	0.2
18〜20	19	4	0.2
20〜22	21	1	0.05

ヒストグラム



1.2　代表値

データ全体の特徴を適当な 1 つの数値で表すことがあり、その数値をデータの**代表値**と言います。よく用いられる代表値として**平均値**、**中央値**（**メジアン**）、**最頻値**（**モード**）があります。まず、平均値ですが、変量 x の平均値を $\bar{x}$ と表し、次のように計算できます。

平均値

$$\bar{x} = \frac{\text{合計}}{\text{データの個数}} = \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_{n-1} + x_n)$$

A地点の気温が、次のように高めで安定しているときに用いる**仮平均法**というものがあります。

28, 29, 27, 30, 26, 31, 32, 25, 33, 30 (℃)

まず、平均に近そうな値や、データの中で最もよく登場する値（最頻値）などを**仮平均**として設定します。例えば**最頻値**（**モード**ともいう）であり、平均とも近そうな 30 を仮平均とすると、各データの仮平均との差は $-2,\ -1,\ -3,\ 0,\ -4,\ 1,\ 2,\ -5,\ 3,\ 0$ この平均は

$$\frac{1}{10}\{(-2) + (-1) + (-3) + 0 + (-4) + 1 + 2 + (-5) + 3 + 0\} = -0.9$$

よって、平均は $30 - 0.9 = 29.1$ です。

仮平均法

仮平均を c として

$\bar{x}$= (仮平均) + (仮平均との差の平均）$= c + \overline{x - c}$

$$\because\ c + \overline{x - c} = c + \frac{1}{n}\{(x_1 - c) + (x_2 - c) + (x_3 - c) + \cdots + (x_n - c)\}$$

$$= c + \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n - nc) = c + \bar{x} - c = \bar{x}$$

平均値は、たった 1 つの外れ値のせいで大きく変化してしまうという弱点があります。例えば 5 人の所持金が 1,000 円, 2,000 円, 3,000 円, 4,000 円, 1,000,000 円だった場合、平均は 202,000 円ですが、これが実態を表しているとは言いづらく、このようなとき平均値はデータを代表する値としてあまり適切とは言えません。そのようなときは、データを小さい順に並べたときの中央の値を代表値とするのが解決策となり得ます。これが**中央値**（**メジアン**）で、上の例の場合は 3,000 円です。データの個数が偶数個になっていて、ちょうど真ん中の値がない場合は、中央に並ぶ二つの値の平均値を中央値とします。

また、上の例での 1,000,000 円のように他の値から極端に離れた値のことを**外れ値**と言います。

中央値

中央値は、$\begin{cases} \text{データの数が奇数個のとき}\quad \text{ちょうど真ん中の値} \\ \text{データの数が偶数個のとき}\quad \text{真ん中の 2 つの値の平均値} \end{cases}$

例) 1000, 2000, 3000, 4000, 1000000 の中央値　3000

1000, 2000, 3000, 4000, 6000, 1000000 の中央値　3500　$\left(\because\ \dfrac{3000+4000}{2}\right)$

[例題 1]

40 人のクラスで数学の試験を行い、男子 25 人の平均は 60 点、女子 15 人の平均は 50 点であった。クラス全体の平均点を求めよ。

解答

$\dfrac{60\times\ 25+50\times\ 15}{40}=\dfrac{2250}{40}=56.25$　よって 56.25 点

男女の人数が異なるので　$\dfrac{60+50}{2}=55$　ではないのに注意します。

[演習 1]

40 人のクラスで模擬試験を行い、全教科の総合点について男子 25 人の平均は 398 点、女子 15 人の平均は 406 点であった。クラス全体の平均点を求めよ。

解答

$398-400=-2$　$406-400=6$

$400+\dfrac{-2\times\ 25+6\times\ 15}{40}=400+1=401$　よって 401 点

400 を仮平均として仮平均法を使いました。

[例題 2]

以下のデータにおいて、x が自然数の値をとって動くとき、中央値は何通り考えられるか。

39, 25, 13, 51, x

解答

$\wedge 13\wedge 25\wedge 39\wedge 51\wedge$　x がどこに入るかで中央値がどうなるかを考える。

データの個数が 5 つであることに注意すると

$\begin{cases} x\ 13\ 25\ 39\ 51 \\ 13\ x\ 25\ 39\ 51 \end{cases}$ のとき、すなわち $x\leqq 25$ のとき中央値は 25

13 25 x 39 51 のとき、すなわち $26 \leqq x \leqq 38$ のとき中央値は x であり、13 通りの値がとれる。

$\begin{cases} 13\ 25\ 39\ x\ 51 \\ 13\ 25\ 39\ 51\ x \end{cases}$ のとき、すなわち $39 \leqq x$ のとき中央値は 39

$\therefore 1 + 13 + 1 = 15$ 通り

[演習 2]

以下のデータにおいて、x が自然数の値を取って動くとき、中央値は何通り考えられるか。

41, 29, 22, 35, 14, x

解答

$\wedge 14 \wedge 22 \wedge 29 \wedge 35 \wedge 41 \wedge$ x がどこに入るかで中央値がどうなるかを考える。

データの個数が 6 つであることに注意すると

$\begin{cases} x\ 14\ 22\ 29\ 35\ 41 \\ 14\ x\ 22\ 29\ 35\ 41 \end{cases}$ のとき、すなわち $x \leqq 22$ のとき中央値は $\dfrac{22+29}{2} = 25.5$

$\begin{cases} 14\ 22\ x\ 29\ 35\ 41 \\ 14\ 22\ 29\ x\ 35\ 41 \end{cases}$ のとき、すなわち $23 \leqq x \leqq 34$ のとき中央値は $\dfrac{x+29}{2}$ であり

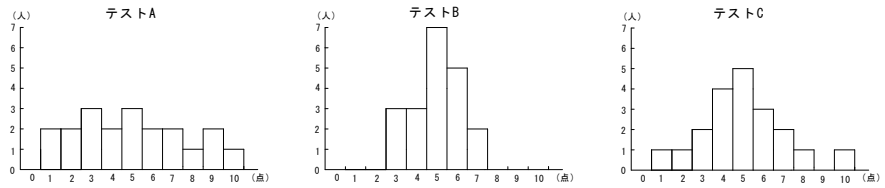
12 通り考えられる。

$\begin{cases} 14\ 22\ 29\ 35\ x\ 41 \\ 14\ 22\ 29\ 35\ 41\ x \end{cases}$ のとき、すなわち $35 \leqq x$ のとき中央値は $\dfrac{29+35}{2} = 32$

$\therefore 1 + 12 + 1 = 14$ 通り

1.3 5 数要約と箱ひげ図

平均値や中央値などの一つの値だけでデータを表すと、違いが分かりにくい場合があります。例えば、下の度数分布表を見てみましょう。



どのデータも、中央値は5点、平均値は5点です。 しかし、データの散らばりの様子にはかなりの違いが見られます。

これらの度数分布表では平均値と中央値がそれぞれ等しくなっています。しかし、データの散らばりについてはずいぶん違います。そこでまず、**最大値**と**最小値**を考えます。こうすることで最大値・中央値・最小値という 3 つの値でデータを見ることになり、特徴をとらえやすくなります。最大値と最小値の差を**範囲**といいます。しかし、上のグラフの A と C を見ると、3 つの値が同じであるのにグラフの様子は違います。そこで**四分位数**というものを考えます。データの値の大きさを小さい順に並べたときに 4 等分する位置にくる値を四分位数といいます。四分位数は小さい方から**第 1 四分位数、第 2 四分位数、第 3 四分位数**と言い、これらを順に Q_1, Q_2, Q_3 で表します。**第 2 四分位数は中央値**です。ただ、ぴったり 4 等分にあたる数字がない場合もあります。そこで、データの値を小さい方から順に左から並べたとき、左半分のデータを下位のデータ、右半分のデータを上位のデータと呼ぶことにした場合、**第 1 四分位数**, **第 3 四分位数をそれぞれ下位のデータの中央値**, **上位のデータの中央値**として定めます。(ただし、データの大きさが奇数のとき、中央の位置にくる値は上位のデータにも下位のデータにも含めないものとします)

下位のデータ　　上位のデータ

2, 3, 5, 7, 11, 13, 18, 20, 23

$Q_1 = \dfrac{3+5}{2} = 4$,　$Q_2 = 11$,　$Q_3 = \dfrac{18+20}{2} = 19$

下位のデータ　　上位のデータ

2, 3, 5, 7, 11, 13, 17, 19, 23, 29

$Q_1 = 5$,　$Q_2 = \dfrac{11+13}{2} = 12$,　$Q_3 = 19$

四分位数

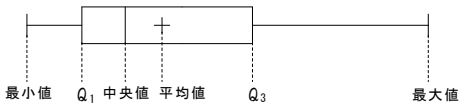
第 1 四分位数 Q_1　下位のデータの中央値　　**第 2 四分位数** Q_2　中央値

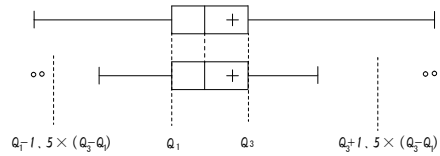
第 3 四分位数 Q_3　上位のデータの中央値

また，$Q_3 - Q_1$ のことを**四分位範囲**，$\dfrac{Q_3 - Q_1}{2}$ のことを**四分位偏差**と言う。

（どちらもデータの散らばりの度合を表す指標であり、これらが大きいほど散らばりの度合が大きいと考えられる）

最小値、第 1 四分位数、中央値、第 3 四分位数、最大値を箱と線（ひげ）で表現する図を**箱ひげ図**と呼び、複数のデータを比較する際に便利です。なお、箱ひげ図に平均値を記入することもあり、その場合、平均値に対応するところに + 印をつけます。また、最小値、第 1 四分位数、中央値、第 3 四分位数、最大値の 5 つの数でデータを表すことを **5 数要約**と言いいます。



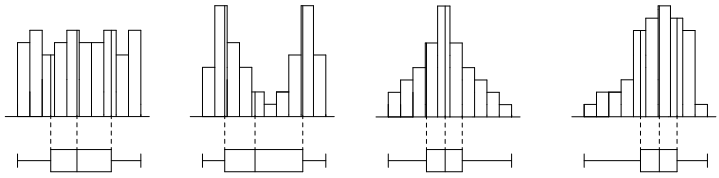


外れ値を含む場合は、通常の左上図ではなく右上図のような箱ひげ図が用いられることがあります。外れ値の基準は複数ありますが、例えば次のような値を外れ値とします。

外れ値

第 1 四分位数 $-$ 1.5 $\times$ 四分位範囲,　第 3 四分位数 $+$ 1.5 $\times$ 四分位範囲

外れ値は○で示しています。また、箱ひげ図の左右のひげはデータから外れ値を除いたときの最小（大）値まで引いています。ただし、四分位数は外れ値を除かないすべてのデータの四分位数であり、その値にもとづいて箱を書きます。(右上図で通常のものと比較しています)
下にはヒストグラムと箱ひげ図の対応例を載せました。箱ひげ図とヒストグラムの対応を見る際には、まず最大値と最小値が合致しているかを見るのが基本です。その後は、ヒストグラムがデータの個数（度数）を面積で表していることを利用して、面積を 4 等分している位置として Q_1, Q_2, Q_3 の目安をつけていきます。ヒストグラムの柱が長いほどその値にデータが密集していることになり、箱ひげ図の幅は小さくなります。



[例題 3]

以下のデータについて、5 数要約、範囲、四分位範囲、四分位偏差を求めよ。

(1) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

(2) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13

(3) 1, 2, 3, 4, $\cdots$, $4m+3$ (m は自然数)

解答

(1)　下位のデータ　　上位のデータ

1, 2, 3, 4, 5, 6,	7, 8, 9, 10, 11, 12

最小値 : 1,　$Q_1 : \dfrac{3+4}{2} = 3.5$,　中央値 : $\dfrac{6+7}{2} = 6.5$,　$Q_3 : \dfrac{9+10}{2} = 9.5$,　最大値 : 12

範囲 : 11,　四分位範囲 : $Q_3 - Q_1 = 6$,　四分位偏差 : $\dfrac{Q_3 - Q_1}{2} = 3$

(2)　下位のデータ　　上位のデータ

1, 2, 3, 4, 5, 6,	7,	8, 9, 10, 11, 12, 13

最小値 : 1,　$Q_1 : \dfrac{3+4}{2} = 3.5$,　中央値 : 7,　$Q_3 : \dfrac{10+11}{2} = 10.5$,　最大値 : 13

範囲 : 12,　四分位範囲 : $Q_3 - Q_1 = 7$,　四分位偏差 : $\dfrac{Q_3 - Q_1}{2} = 3.5$

(3)　下位のデータ　　上位のデータ

$1,\ 2,\ \cdots,\ m+1,\ \cdots,\ 2m+1,$	$2m+2,$	$2m+3,\ \cdots,\ 3m+3,\ \cdots,\ 4m+3$

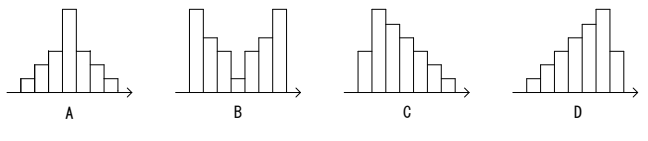
最小値 : 1,　$Q_1 : m+1$,　中央値 : $2m+2$,　$Q_3 : 3m+3$,　最大値 : $4m+3$

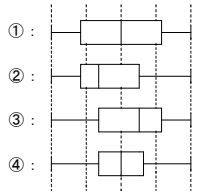
範囲 : $4m+2$,　四分位範囲 : $Q_3 - Q_1 = 2m+2$,　四分位偏差 : $\dfrac{Q_3 - Q_1}{2} = m+1$

結局、データの個数を 4 で割った余りが 0, 1, 2, 3 の場合で状況が異なっています。

[演習 3]

次のヒストグラムが表すデータと対応する箱ひげ図はどれか答えよ。





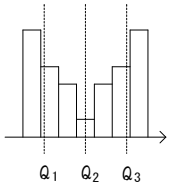
解答

A − ④　B − ①　C − ②　D − ③

最大値と最小値はすべてのグラフで一致しているので違う部分に注目します。まず、左右対称なのは A と B です。両者とも中央値はど真ん中にありますが、A は中央付近の柱が長くなっているので、Q_1〜Q_3 の間を表す箱の幅が小さいと分かります。よって、A は ④ です。つまり B は ① となります。

Q_1, Q_3 の位置は結局 4 等分の位置ですから、最小値と中央値の間の面積を半分に分けるとこ

ろに線を引けば、Q_1 の目星が付きます。



残るはCとDですが、面積を2等分するところが中央値であるので、Cは左寄り、Dは右寄りにあることが分かります。よって、Cは ②、Dは ③ です。

1.4　分散，標準偏差

平均値からの散らばり具合を数値でとらえることを考えます。まず、各値と平均との差の合計を考えてみます。（各値と平均との差のことを**平均値からの偏差**または単に**偏差**といいます）すると　$(x_1-\bar{x})+(x_2-\bar{x})+(x_3-\bar{x})+\cdots+(x_n-\bar{x})=x_1+x_2+\cdots+x_n-n\bar{x}=0$
平均とはそうなるように設定したのだから当然です。しかし、これでは散らばり具合が分かりません。そこで、一つの方法として偏差を2乗してみます。
$(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+\cdots+(x_n-\bar{x})^2$
しかし、偏差の2乗の和を考えただけではデータが多くなればなるほど和が大きくなるためデータの個数が異なるデータの散らばり具合を比較できません。よって、偏差の2乗の和をデータの個数で割った値、すなわち**偏差の2乗の平均値**を考えます。これを**分散**と呼び、s^2 で表します。$s^2=\dfrac{1}{n}\{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+\cdots+(x_n-\bar{x})^2\}$
分散は散らばりを表す値として適切ですが、単位が元の値と変わってしまっているという問題があります。$\sqrt{\text{分散}}$ を考えると、この問題が解決され、これを**標準偏差**と呼び、s で表します。

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

ここで、分散の定義の式を展開して整理してみましょう。

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

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

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

$$=\overline{x^2}-2\bar{x}\bar{x}+(\bar{x})^2=\overline{x^2}-(\bar{x})^2$$

すなわち、分散は　$s^2=(\text{2 乗の平均})-(\text{平均の 2 乗})$　とも表せることが分かります。

分散，標準偏差

分散　$s^2 =$ **偏差の 2 乗の平均値** $= \dfrac{1}{n}\{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+\cdots+(x_n-\bar{x})^2\}$

分散の別公式　$s^2 = (2$ **乗の平均**$) - ($**平均の 2 乗**$) = \overline{x^2} - (\bar{x})^2$

標準偏差　$s = \sqrt{\text{**分散**}} = \sqrt{\dfrac{1}{n}\{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+\cdots+(x_n-\bar{x})^2\}}$

ところで、分散の別公式は (**2 乗の平均**)−(**平均の 2 乗**) ですが、どちらからどちらを引くのかを忘れがちです。しかし、分散が必ず 0 以上の値であることを確認し、その上で平均 0 の周りに正負に散らばっているデータ (1, 2, 5, −2, −3, −3) をイメージすれば間違えないでしょう。

[例題 4]

次のデータの分散と標準偏差を求めよ。

(1) 16, 26, 16, 26, 31

(2) 3, 6, 4, 6, 2

解答

(1)

	x	$x-\bar{x}$	$(x-\bar{x})^2$
	16	−7	49
	26	3	9
	16	−7	49
	26	3	9
	31	8	64
合計	115	0	180
平均	23	0	36

表より　$s^2 = 36$　よって　$s = 6$

表を書いて求めるのが良いでしょう。$x-\bar{x}$ の合計や $x-\bar{x}$ の平均が 0 となることを必ず確認しましょう。

(2)

	x	x^2
	3	9
	6	36
	4	16
	6	36
	2	4
合計	21	101
平均	4.2	20.2

表より $s^2 = 20.2 - 4.2^2 = 20.2 - 17.64 = 2.56$ よって $s = 1.6$

このように平均が小数になった場合、偏差の 2 乗の計算が面倒です。このような場合は、分散の別公式が有効です。

[演習 4]
25 個の値からなるデータがあり、そのうちの 10 個の値の平均値は 4、分散は 14、残りの 15 個の値の平均値は 9、分散は 19 である。このデータの平均値と分散を求めよ。

解答

25 個のデータのうち 10 個のデータの方を x、15 個のデータの方を y とする。

まず (全体の平均) $= \dfrac{4\times\ 10 + 9\times\ 15}{25} = 7$

また $s_x{}^2 = \overline{x^2} - 4^2 = 14 \quad \therefore\ \overline{x^2} = 30$ よって、x^2 の合計は 300

$s_y{}^2 = \overline{y^2} - 9^2 = 19 \quad \therefore\ \overline{y^2} = 100$ よって、y^2 の合計は 1500

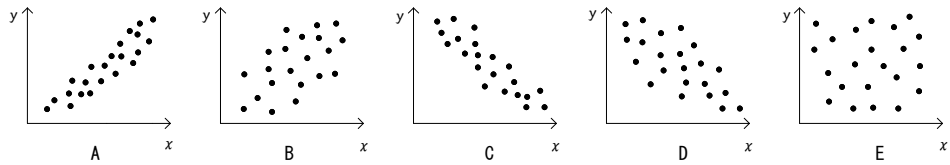
よって、(全体の分散) $= \dfrac{1}{25}(300 + 1500) - 7^2 = 72 - 49 = 23$

2 つのデータを合わせたときの平均と分散を求める問題です。まず、各データの値が分からないので分散の別公式に頼るしかありません。また、分散は平均のように $\dfrac{14\times 10 + 19\times 15}{25}$ という計算では求まらないことに注意します。2 つのデータを合わせると平均値が変わってしまうので、平均値周りの散らばり具合である分散も影響を受けます。(本問では、合わせた分散がどちらの分散よりも大きくなりました)

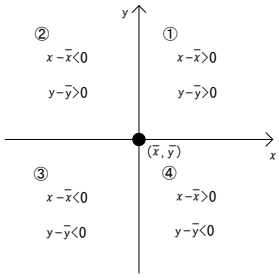
1.5 共分散・相関係数

2 つの変量の関係について調べたいとき、例えば下の表にある数学の成績 (x) と物理の成績 (y) の関係を (x, y) を座標としてグラフにした図を描きます。このような図を**散布図**と言います。下図 A, B の場合、x が増えるに連れて y も増えるという傾向が見てとれ、そのようなとき

「正の相関関係がある」と言います。（直線により近い A の方が B より「強い正の相関関係がある」と言えます）その他に、x が増えるに連れて y が減る傾向があるとき（下図 C, D）、**「負の相関関係がある」**と言い、どちらの関係も認められないとき（下図 E)、**「相関関係がない」**と言います。



しかしながら、相関関係の有無を見た目だけで判断するのでは不十分です。そこで、目安となる数値を考えてみましょう。2 つの変量 $(x,\ y)$ のデータが次のように n 組与えられているとします。$(x_1,\ y_1)$, $(x_2,\ y_2)$, $(x_3,\ y_3)$, $\cdots$, $(x_n,\ y_n)$ このとき $x,\ y$ の平均値 $\overline{x},\ \overline{y}$ をもとに、座標平面を 4 分割することを考え、上記の n 組を点としてプロットしていきます。



そうすると、①, ③の領域に点が多くあるほど正の相関関係があると考えらえれ、②, ④に多くあるほど負の相関関係があると考えられます。ところで、①, ③の領域において偏差の積 $(x-\overline{x})(y-\overline{y})$ は正となり、②, ④の領域においては偏差の積は負となることが分かります。よって

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

を考えて、全体として値が正になっていれば、①, ③の領域により多くの点がプロットされていると考えられ（正の相関関係がある）、負になっていれば②, ④の領域により多くの点がプロットされていると考えられます（負の相関関係がある）。よって、偏差の積の和は相関の見当をつけるのには適切な値ですが、データが多くなればなるほど値が大きくなるので、分散のときのように平均値を考えます。すなわち**偏差の積の平均値**で、これが**共分散**であり、s_{xy} で表します。

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

共分散は、分散と同様に別公式があります。展開してみると

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

$$= \frac{1}{n}\{(x_1y_1+x_2y_2+x_3y_3+\cdots+x_ny_n)-(x_1+x_2+x_3+\cdots+x_n)\overline{y}-(y_1+y_2+y_3+\cdots+y_n)\overline{x}+n\overline{xy}\}$$

$$= \frac{1}{n}(x_1y_1+x_2y_2+x_3y_3+\cdots+x_ny_n)-\frac{1}{n}(x_1+x_2+x_3+\cdots+x_n)\overline{y}-\frac{1}{n}(y_1+y_2+y_3+\cdots+y_n)\overline{x}+\overline{xy}$$

$$= \overline{xy} - \overline{x}\,\overline{y} - \overline{y}\,\overline{x} + \overline{xy}$$

$$= \overline{xy} - \overline{x}\,\overline{y}$$

すなわち、共分散は $s_{xy} = (\textbf{積の平均}) - (\textbf{平均の積})$ とも表せることが分かりました。

共分散

共分散 $s_{xy} = \textbf{偏差の積の平均値}$

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

共分散の別公式 $s_{xy} = (\textbf{積の平均}) - (\textbf{平均の積}) = \overline{xy} - \bar{x}\bar{y}$

共分散はその正負でデータの相関が分かる有用な数値です。しかし、欠点もあります。データの個数が一緒で、本質的にデータの散らばりに違いがないような「(1, 2) (2, 4) (3, 6) (4, 8) (5, 10)」と「(10, 20) (20, 40) (30, 60) (40, 80) (50, 100)」を考えたとき、後者の共分散が前者のそれの 100 倍になってしまいます。これを解決するために、共分散を標準偏差の積で割ったものを考えます。それを**相関係数**と呼びます。

相関係数 $r = \dfrac{s_{xy}}{s_x \times s_y} = \dfrac{\textbf{共分散}}{(x\ \textbf{の標準偏差}) \times (y\ \textbf{の標準偏差})}$

相関係数にはいくつかの性質があり、大体どれぐらいの値のときにどのような図になるかも知っておくべきです。

相関係数

相関係数 $r = \dfrac{s_{xy}}{s_x \cdot s_y} = \dfrac{\text{共分散}}{(x\ \text{の標準偏差}) \times (y\ \text{の標準偏差})}$

$$= \frac{\frac{1}{n}\{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})\}}{\sqrt{\frac{1}{n}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\}}\sqrt{\frac{1}{n}\{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2\}}}$$

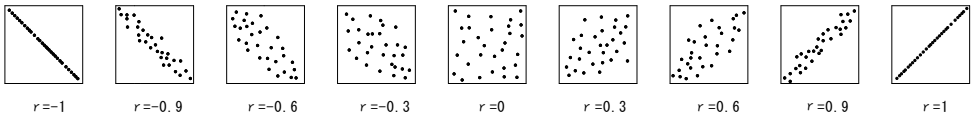
相関係数の性質

(i) $-1 \leqq r \leqq 1$

(ii) $r = 1$ のとき右上がりの直線に沿って分布。1 に近いほど強い正の相関関係がある。

(iii) $r = -1$ のとき右下がりの直線に沿って分布。-1 に近いほど強い負の相関関係がある。

(iv) r の値が 0 に近いとき相関関係がない。



$(x_1,\ y_1),\ (x_2,\ y_2),\ (x_3,\ y_3),\ \cdots,\ (x_n,\ y_n)$ の相関係数が $-1 \leqq r \leqq 1$ となる理由を考えてみましょう。まず一般に、コーシーシュワルツの不等式は n 項の場合にも成立します。

$$(x_1{}^2 + x_2{}^2 + \cdots + x_n{}^2)(y_1{}^2 + y_2{}^2 + \cdots + y_n{}^2) \geqq (x_1y_1 + x_2y_2 + \cdots + x_ny_n)^2$$

$$\therefore\ 1 \geqq \frac{(x_1y_1 + x_2y_2 + \cdots + x_ny_n)^2}{(x_1{}^2 + x_2{}^2 + \cdots + x_n{}^2)(y_1{}^2 + y_2{}^2 + \cdots + y_n{}^2)} \quad \cdots ①$$

ここで $X_i = x_i - \bar{x},\ Y_i = y_i - \bar{y}$ とします。このとき

$$s_x{}^2 = \frac{1}{n}\{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\} = \frac{1}{n}(X_1{}^2 + X_2{}^2 + \cdots + X_n{}^2)$$

$$s_y{}^2 = \frac{1}{n}\{(y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + \cdots + (y_n - \overline{y})^2\} = \frac{1}{n}(Y_1{}^2 + Y_2{}^2 + \cdots + Y_n{}^2)$$

$$s_{xy} = \frac{1}{n}\{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \cdots + (x_n - \overline{x})(y_n - \overline{y})\} = \frac{1}{n}(X_1Y_1 + X_2Y_2 + \cdots + X_nY_n)$$

よって $r_{xy} = \dfrac{s_{xy}}{s_x \cdot s_y} = \dfrac{s_{xy}}{\sqrt{s_x{}^2}\sqrt{s_y{}^2}} = \dfrac{\frac{1}{n}(X_1Y_1 + X_2Y_2 + \cdots + X_nY_n)}{\sqrt{\frac{1}{n}(X_1{}^2 + X_2{}^2 + \cdots + X_n{}^2)}\sqrt{\frac{1}{n}(Y_1{}^2 + Y_2{}^2 + \cdots + Y_n{}^2)}}$

$$= \frac{(X_1Y_1 + X_2Y_2 + \cdots + X_nY_n)}{\sqrt{(X_1{}^2 + X_2{}^2 + \cdots + X_n{}^2)}\sqrt{(Y_1{}^2 + Y_2{}^2 + \cdots + Y_n{}^2)}}$$

$$\therefore\ r_{xy}{}^2 = \frac{(X_1Y_1 + X_2Y_2 + \cdots + X_nY_n)^2}{(X_1{}^2 + X_2{}^2 + \cdots + X_n{}^2)(Y_1{}^2 + Y_2{}^2 + \cdots + Y_n{}^2)}$$

よって、① より $1 \geqq r_{xy}{}^2 \Longleftrightarrow -1 \leqq r_{xy} \leqq 1$ となり、相関係数が -1〜1 の間にあることが分かりました。

二つの変量の間の関係を調べるときに、2 つの度数分布表を組み合わせた相関表と呼ばれる表を用いることがあります。以下のようなもので、データの値の組が多くて、散布図だと点がいくつも重なって、とらえにくいときに有用です。

x(cm) / y(kg)	170 以上 175 未満	175〜180	180〜185	185〜190	計
90 以上 100 未満		1	1	2	4
80〜90		4	6		10
70〜80	7	4	4		15
60〜70	1				1
計	8	9	11	2	30

[例題 5]

次の変量 $(x,\ y)$ に対する相関係数を求めよ。

$(4, 6), (3, 8), (2, 12), (6, 10), (5, 4)$

解答

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
	4	6	0	-2	0	4	0
	3	8	-1	0	1	0	0
	2	12	-2	4	4	16	-8
	6	10	2	2	4	4	4
	5	4	1	-4	1	16	-4
合計	20	40	0	0	10	40	-8
平均	4	8	0	0	2	8	-1.6

表より $r_{xy} = \dfrac{s_{xy}}{s_x \cdot s_y} = \dfrac{-1.6}{\sqrt{2}\sqrt{8}} = \dfrac{-1.6}{4} = -0.4$

相関係数を求めるには、分散や標準偏差、共分散を求めなければならず、ばらばらに計算するとミスしやすいので表などを利用します。

[演習 5]

次の変量 $(x,\ y)$ に対する相関係数を求めよ。$(1,\ 3), (3,\ 4), (5,\ 5), (9,\ 7), (11,\ 8)$

解答

	x	y	x^2	y^2	xy
	1	3	1	9	3
	3	4	9	16	12
	5	5	25	25	25
	9	7	81	49	63
	11	8	121	64	88
合計	29	27	237	163	191
平均	5.8	5.4	47.4	32.6	38.2

表より $s_{xy} = \overline{xy} - \overline{x}\overline{y} = 38.2 - 5.8 \times 5.4 = 6.88$

${s_x}^2 = \overline{x^2} - \overline{x}^2 = 47.4 - 5.8^2 = 13.76$

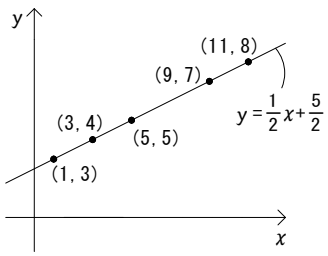
${s_y}^2 = \overline{y^2} - \overline{y}^2 = 32.6 - 5.4^2 = 3.44$

$\therefore\ r_{xy} = \dfrac{6.88}{\sqrt{13.76}\sqrt{3.44}} = \dfrac{688}{\sqrt{1376}\sqrt{344}} = 1$

例題のように分散や共分散を計算しようとして偏差を考えると、$\overline{x}$ や $\overline{y}$ が小数になっているため、かなり大変です。このようなときは分散や共分散の別公式が有効です。

また、本問では相関係数が 1 になりました。$(x,\ y)$ の散布図を描いてみると図のようになっており、$y = \dfrac{1}{2}x + \dfrac{5}{2}$ という直線上にすべての点があることが分かります。

結局、傾きが正の直線上にすべての点があれば相関係数 1 ということです（負の直線上なら -1 で、相関係数の 1 という数字はこの直線の傾きや切片などには関係ありません）



1.6　変量の変換

データの各値に一斉に同じ数を加えたり掛けたりしたときに（**変量の変換**）、平均値・分散・標準偏差がどうなるのかを考えます。変量 x について $p = ax + b$ で新たな変量 p を考えます。そうすると

$$\overline{p} = \frac{1}{n}(p_1 + p_2 + p_3 + \cdots + p_n) = \frac{1}{n}\{(ax_1 + b) + (ax_2 + b) + \cdots + (ax_n + b)\}$$

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

ここで、p の偏差を考えると

$p_k - \overline{p} = ax_k + b - (a\overline{x} + b) = a(x_k - \overline{x})$ であることより

$$s_p{}^2 = \frac{1}{n}\{(p_1 - \overline{p})^2 + (p_2 - \overline{p})^2 + \cdots + (p_n - \overline{p})^2\} = \frac{1}{n}\{a^2(x_1 - \overline{x})^2 + a^2(x_2 - \overline{x})^2 + \cdots + a^2(x_n - \overline{x})^2\}$$

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

$$s_p = \sqrt{a^2 s_x{}^2} = |a| s_x$$

となります。**データの各値に b を加えると、平均値も b だけ増加します。**一方で、**分散や標準偏差は**平均値周りの散らばり具合ですから、**全体が b 増加しても変わりません。**また、**全体に a を掛けた場合は、**データの各値も**平均値も a 倍になる**ので、各値と平均値との偏差も a 倍となり、**分散が a^2 倍、標準偏差が $|a|$ 倍になります。**意味で捉えて覚えておきましょう。

変量の変換による平均値, 分散, 標準偏差の変化

元の変量	平均値	分散	標準偏差
b 加える	b 加わる	変化なし	変化なし
a 倍する	a 倍になる	a^2 倍になる	$\|a\|$ 倍になる

[例題 6]

5 人のデータ身長 $x\,(cm)$ のデータ

176, 170, 167, 179, 168

に対して $x_0 = 170$ として、新たな変量 u を $u = x - x_0$ で定める。

変量 u の平均と分散を求めよ。また、変量 x の平均と分散は変量 u の平均と分散からどのように変化しているか述べよ。

解答

$$\overline{u} = \frac{1}{5}\{6 + 0 + (-3) + 9 + (-2)\} = 2$$

$$s_u{}^2 = \overline{u^2} - (\overline{u})^2 = \frac{1}{5}(36 + 0 + 9 + 81 + 4) - 2^2 = 26 - 4 = 22$$

$x = u + x_0$ より $\overline{x} = \overline{u} + x_0 = 2 + 170 = 172$ x の平均は u の平均に比べて 170 増えている。

${s_x}^2 = {s_u}^2 = 22$　x の分散は u の分散と変化なし。

仮平均を 170 としたときの仮平均との差の平均、分散を求めているわけです。

仮平均の公式からも明らかですが、x は u のそれぞれのデータに 170 を足したものですから、平均は 170 増えることになります。

一方で、分散は平均値まわりの散らばり具合のことなので、全データに 170 を足した場合、平均も 170 も増えるので、そこからの散らばり具合には変化がありません。

[演習 6A]

変量 x のデータが次のように与えられている。

750, 740, 720, 770, 750, 740

今　$c = 10,\ x_0 = 740,\ u = \dfrac{x - x_0}{c}$　として新たな変量 u をつくる。

(1) 変量 u のデータの平均値と標準偏差を求めよ。

(2) 変量 x のデータの平均値と標準偏差を求めよ。

解答

(1) u は $\dfrac{10}{10},\ \dfrac{0}{10},\ \dfrac{-20}{10},\ \dfrac{30}{10},\ \dfrac{10}{10},\ \dfrac{0}{10}$ つまり $1,\ 0,\ -2,\ 3,\ 1,\ 0$

よって　$\overline{u} = \dfrac{1}{6}\{1 + 0 + (-2) + 3 + 1 + 0\} = \dfrac{1}{2}$

${s_u}^2 = \overline{u^2} - (\overline{u})^2 = \dfrac{1}{6}(1 + 0 + 4 + 9 + 1 + 0) - \left(\dfrac{1}{2}\right)^2 = \dfrac{9}{4}$　$\therefore\ s_u = \sqrt{\dfrac{9}{4}} = \dfrac{3}{2}$

別解

（u, $\overline{u}$を求めるところまで同じ）

${s_u}^2 = \dfrac{1}{6}\left\{\left(1 - \dfrac{1}{2}\right)^2 + \left(0 - \dfrac{1}{2}\right)^2 + \left(-2 - \dfrac{1}{2}\right)^2 + \left(3 - \dfrac{1}{2}\right)^2 + \left(1 - \dfrac{1}{2}\right)^2 + \left(0 - \dfrac{1}{2}\right)^2\right\}$

$= \dfrac{1}{6}\left(\dfrac{1}{4} + \dfrac{1}{4} + \dfrac{25}{4} + \dfrac{25}{4} + \dfrac{1}{4} + \dfrac{1}{4}\right) = \dfrac{9}{4}$　$\therefore s_u = \sqrt{\dfrac{9}{4}} = \dfrac{3}{2}$

(2) $x = cu + x_0$　$\therefore \bar{x} = c\bar{u} + x_0 = 10 \times \dfrac{1}{2} + 740 = 745$

また，$s_x = |c| s_u = 10 \times \dfrac{3}{2} = 15$

仮平均 740 を定めて、さらに 10 で割ることで、随分とデータが扱いやすくなりました。

[演習 6B]

試験の得点 x 点に対して偏差値 T は次の式で計算される。

$T = 10 \times \dfrac{x - \bar{x}}{s_x} + 50$

(1) 偏差値 T の平均値と標準偏差を求めよ。

(2) 45 人が試験を受けて結果が以下のようになった。テストの得点の平均値、標準偏差を求めよ。また、偏差値が 60 以上の生徒の人数をそれぞれ求めよ。

得点	0 点	10 点	20 点	30 点	40 点	50 点	60 点	70 点
人数	1	3	5	9	13	7	5	2

解答

(1) $T = \dfrac{10}{s_x}x - \dfrac{10}{s_x}\bar{x} + 50 \quad \therefore\ \overline{T} = \dfrac{10}{s_x}\bar{x} - \dfrac{10}{s_x}\bar{x} + 50 = 50, \quad s_T = \left|\dfrac{10}{s_x}\right| s_x = \dfrac{10}{s_x}s_x = 10$

$x - \bar{x}$ の平均は 0 なので、$10 \times \dfrac{x - \bar{x}}{s_x}$ の部分の平均が 0 となり、その部分を除いて平均は 50 と考えても構いません。偏差値の平均は常に 50 で標準偏差は常に 10 ということです。

(2) $\bar{x} = 40 + \dfrac{1}{45}\{(-40)\cdot 1 + (-30)\cdot 3 + (-20)\cdot 5 + (-10)\cdot 9 + 0\cdot 13 + 10\cdot 7 + 20\cdot 5 + 30\cdot 2\} = 38$

また $X = x - 40$ とすると $\overline{X} = \bar{x} - 40 = -2$

よって ${s_x}^2 = {s_X}^2 = \overline{X^2} - (\overline{X})^2 = \dfrac{1}{45}\{(-40)^2\cdot 1 + (-30)^2\cdot 3 + (-20)^2\cdot 5 + (-10)^2\cdot 9$

$+ 0^2\cdot 13 + 10^2\cdot 7 + 20^2\cdot 5 + 30^2\cdot 2\} - (-2)^2 = \dfrac{1}{45}(1600 + 2700 + 2000 + 900 + 700$

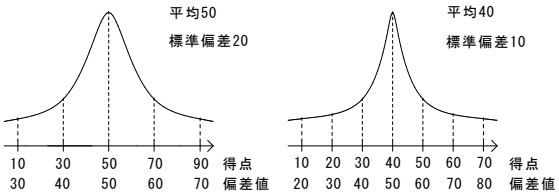
$+ 2000 + 1800) - 4 = 260 - 4 = 256 \quad \therefore\ s_x = 16$

$10\cdot\dfrac{x - 38}{16} + 50 \geqq 60 \Longleftrightarrow x \geqq 54$ よって、偏差値 60 以上の人数は 54 点以上の 7 人。

最頻値である 40 を仮平均として計算しました。${s_x}^2$ は $\overline{x^2} - (\bar{x})^2$ として計算してもよいのですが、仮平均を設定して変量の変換の知識を使って計算をすると楽に求めることができる場合があります。

結局、平均点のとき偏差値が 50 となり、偏差値の定義式から偏差値を 10 あげるには標準偏差の分 (本問だと 16 点) だけ点数をあげる必要があるということです。

ところで、本問では偏差値 60 以上の生徒は 45 人中 7 人だったので、全体の約 15.6%が偏差値 60 以上だったことになります。一般に、母集団のデータが正規分布（図のように左右対称で平均値付近にデータが多数あるような分布）に従うとき、偏差値 70 以上は約 2.3%、60 以上は約 16%、50 以上は約 50%、40 以下は約 16%、30 以下は約 2.3%が分布します。



偏差値は、平均を 50 に標準偏差を 10 に換算する効果があり、平均点や散らばり具合が異なるテストにおいても集団のおおよそどの辺りに位置しているかが把握できます。例えば、標準偏差の小さい分布の方が同じ平均点+10 点でもより上位に位置することになり、偏差値が高くなります。

次に、共分散 s_{xy} と相関係数 r_{xy} についても考察します。$p=ax+b,\ \ q=cy+d$ で新たな変量 $p,\ q$ を考えます。まず、p と同様にして $\overline{q}=c\overline{y}+d,\ \ {s_q}^2=c^2{s_y}^2,\ \ s_q=|c|s_y$

よって $s_{pq}=\dfrac{1}{n}\{(p_1-\overline{p})(q_1-\overline{q})+(p_2-\overline{p})(q_2-\overline{q})+\cdots+(p_n-\overline{p})(q_n-\overline{q})\}$

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

$=ac\cdot\dfrac{1}{n}\{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})\}=acs_{xy}$

結局、**共分散は元の ac 倍になります。**

また $r_{pq}=\dfrac{s_{pq}}{s_ps_q}=\dfrac{acs_{xy}}{|a|s_x|c|s_y}=\dfrac{ac}{|ac|}\dfrac{s_{xy}}{s_xs_y}=\begin{cases}ac>0\text{ のとき} & \dfrac{s_{xy}}{s_xs_y}\quad(r_{pq}=r_{xy})\\ ac<0\text{ のとき} & -\dfrac{s_{xy}}{s_xs_y}\quad(r_{pq}=-r_{xy}0)\end{cases}$

$x,\ y$ の一方に負の数をかけると相関の正負が逆転するものの、相関係数の絶対値については定数をかけたり足したりしても変化しないということで、面白い性質です。

変量の変換による共分散、相関係数の変化

変量	x	y	$p=ax+b$	$q=cy+d$
平均値	$\bar{x}$	$\bar{y}$	$\bar{p}=a\bar{x}+b$	$\bar{q}=c\bar{y}+d$
分散	${s_x}^2$	${s_y}^2$	${s_p}^2=a^2{s_x}^2$	${s_q}^2=c^2{s_y}^2$
標準偏差	s_x	s_y	$s_p=\|a\|s_x$	$s_q=\|c\|s_y$
共分散	s_{xy}		$s_{pq}=acs_{xy}$	
相関係数	r_{xy}		$r_{pq}=r_{xy}\ (ac>0\text{ のとき})\ \ r_{pq}=-r_{xy}\ (ac<0\text{ のとき})$	

[例題 7]

変量 $x,\ y$ について、それぞれ平均値は 5, 4、分散は 4, 16、共分散は 6.8 であるとする。$p=2x+1,\ \ q=-y+3$ を定めるとき、$p,\ q$ の平均値・分散・標準偏差・共分散・相関係数を求めよ。

解答

まず $\ s_x=\sqrt{4}=2,\ \ s_y=\sqrt{16}=4,\ \ r_{xy}=\dfrac{s_{xy}}{s_x\cdot s_y}=\dfrac{6.8}{2\cdot 4}=0.85$

$\overline{p}=\overline{2x+1}=2\overline{x}+1=2\cdot 5+1=11,\ \overline{q}=\overline{-y+3}=-\overline{y}+3=-4+3=-1$

${s_p}^2=2^2{s_x}^2=16,\ \ {s_q}^2=(-1)^2{s_y}^2=(-1)^2\times\ 16=16$

$s_p=|\,2\,|s_x=2\cdot 2=4,\ \ s_q=|-1|s_y=|-1|\cdot 4=4$

$s_{pq}=2\cdot(-1)s_{xy}=-2\cdot 6.8=-13.6$

$2\cdot(-1)=-2<0$ より，$r_{pq}=-r_{xy}=-0.85$

$s_p=\sqrt{{s_p}^2}=\sqrt{16}=4,\ \ s_q=\sqrt{{s_q}^2}=\sqrt{16}=4,\ \ r_{pq}=\dfrac{s_{pq}}{s_ps_q}=\dfrac{-13.6}{4\cdot 4}=-0.85$ として計算することもできます。変量の具体的な数値が与えられていないので、$p=2x+1,\ q=-y+3$ という関係からそれぞれの値がどのように変化するかを計算しなければなりません。

[演習 7]

(1) 2 つの変量 $x,\ y$ の n 組の値からなるデータ

$(x_1,\ y_1),\ (x_2,\ y_2),\ (x_3,\ y_3),\ \cdots,\ (x_n,\ y_n)$

がある。変量 $x,\ y$ の平均値をそれぞれ $\overline{x},\ \overline{y}$ とし、分散をそれぞれ ${s_x}^2,\ {s_y}^2$ とする。ただし $\ {s_x}^2\neq 0\ $ とする。また、x と y の共分散を s_{xy}、相関係数を r_{xy} とする。

実数 a に対して $\ f(x)=a(x-\overline{x})+\overline{y}\ $ とするとき

$L(a)=\dfrac{1}{n}\left[\{y_1-f(x_1)\}^2+\{y_2-f(x_2)\}^2+\cdots+\{y_n-f(x_n)\}^2\right]$

が最小となるような a の値を ${s_x}^2,\ s_{xy}$ を用いて表せ。また、そのときの $L(a)$ の値を ${s_y}^2$ と r_{xy} を用いて表せ。

(2) (1) で求めた a に対して直線 $\ y=f(x)\ $ を y の x への回帰直線という。変量 $x,\ y$ の 10 組の値からなるデータ

(2, 2), (3, 5), (5, 8), (8, 4), (10, 5), (10, 6), (13, 5), (14, 12), (17, 8), (18, 15)

に対し、回帰直線の方程式を求めよ。

(3) (2) の 10 組のデータ $(x,\ y)$ に元に $\ X=2x,\ \ Y=-y+5\ $ で定める $(X,\ Y)$ に対し、回帰直線の方程式を求めよ。また、このときの $L(a)$ の最小値は (2) の $L(a)$ の最小値の何倍か。

解答

(1) $y_i - f(x_i) = y_i - \{a(x_i - \overline{x}) + \overline{y}\} = y_i - \overline{y} - a(x_i - \overline{x})$

よって $L(a) = \dfrac{1}{n}\left[\{y_1 - \overline{y} - a(x_1 - \overline{x})\}^2 + \{y_2 - \overline{y} - a(x_2 - \overline{x})\}^2 + \cdots + \{y_n - \overline{y} - a(x_n - \overline{x})\}^2\right]$

$= \dfrac{1}{n}\left[\{(y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + \cdots + (y_n - \overline{y})^2\} + a^2\{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\}\right.$

$\left. - 2a\{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \cdots + (x_n - \overline{x})(y_n - \overline{y})\}\right] = {s_y}^2 + a^2{s_x}^2 - 2as_{xy}$

$= {s_x}^2\left(a^2 - \dfrac{2s_{xy}}{{s_x}^2}\,a\right) + {s_y}^2 = {s_x}^2\left(a - \dfrac{s_{xy}}{{s_x}^2}\right)^2 - \dfrac{{s_{xy}}^2}{{s_x}^2} + {s_y}^2$

よって $a = \dfrac{s_{xy}}{{s_x}^2}$ のとき最小となる。このとき $L(a) = {s_y}^2\left(1 - \dfrac{{s_{xy}}^2}{{s_x}^2{s_y}^2}\right) = {s_y}^2\left(1 - {r_{xy}}^2\right)$

(2)

	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
	2	2	-8	-5	64	40
	3	5	-7	-2	49	14
	5	8	-5	1	25	-5
	8	4	-2	-3	4	6
	10	5	0	-2	0	0
	10	6	0	-1	0	0
	13	5	3	-2	9	-6
	14	12	4	5	16	20
	17	8	7	1	49	7
	18	15	8	8	64	64
合計	100	70	0	0	280	140
平均	10	7	0	0	28	14

表より ${s_x}^2 = 28,\ s_{xy} = 14$

よって $a = \dfrac{s_{xy}}{{s_x}^2} = \dfrac{14}{28} = \dfrac{1}{2}$ のとき $L(a)$ は最小となる。

このとき $f(x) = \dfrac{1}{2}(x - 10) + 7 = \dfrac{1}{2}x + 2$

よって $y = \dfrac{1}{2}\,x + 2$

(3) $X = 2x,\ Y = -y + 5$ より $\overline{X} = 2\bar{x} = 20,\ \overline{Y} = -\overline{y} + 5 = -2$

また ${s_X}^2 = 4{s_x}^2,\ {s_Y}^2 = {s_y}^2,\ s_{XY} = -2s_{xy},\ r_{XY} = -r_{xy}$

よって $a = \dfrac{s_{XY}}{{s_X}^2} = \dfrac{-2s_{xy}}{4{s_x}^2} = -\dfrac{1}{4}$

$\therefore\ f(x) = -\dfrac{1}{4}(x-20) - 2 = -\dfrac{1}{4}x + 3$

$\therefore\ y = -\dfrac{1}{4}x + 3$

また $L(a) = {s_Y}^2(1-{r_{XY}}^2) = {s_y}^2(1-{r_{xy}}^2)$ である。 $\therefore\ L(a)$ は (2) の $L(a)$ の 1 倍。

ところで、(2) のデータの散布図に求めた直線を重ねると図のようになります。$L(a)$ はデータの各値と直線との y 方向の距離 $|y_i - f(x_i)|$ の 2 乗の平均値であり、この値が小さいほど「**回帰直線**が $x,\ y$ の関係をよく表している」と言えます。このように $|y_i - f(x_i)|$ の 2 乗の平均値が最小になるように $y = f(x)$ を定める方法を**最小 2 乗法**といいます。また、$L(a)$ の最小値) $= {s_y}^2\left(1-{r_{xy}}^2\right)$ と表せるので、$r_{xy} = \pm 1$ **のとき、最小値が 0 すなわちデータを構成するすべての** $(x,\ y)$ **の組が回帰直線上にある**ことが分かります。