数学Ⅰ

第５章

データの分析

▌1　データの分析

I.　度数分布表とヒストグラム

階級：区切られた区間

階級の幅：区間の幅

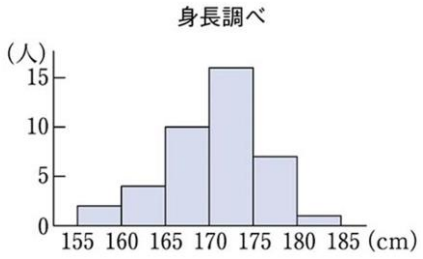
度数：階級に含まれる値の個数

階級値：各階級の中央の値

度数分布表を柱状グラフに表したものがヒストグラムである。左下の度数分布表をヒストグラムに表すと右下の図のようになる。

身長調べ

階級(cm)	度数(人)
155以上160未満	2
160 ～ 165	4
165 ～ 170	10
170 ～ 175	16
175 ～ 180	7
180 ～ 185	1
計	40



II.　データの分析

①　範囲R

データの最大値Mから最小値mを引いたものを範囲Rという。

$$R = M - m$$

②　代表値

(1) 平均値$\overline{x}$

変量xのデータの値が、$x_1, x_2, \cdots, x_n$であるとき、このデータの平均値$\overline{x}$は

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

(2) 中央値（メジアン）

データを値の大きさの順に並べたとき、中央の位置にくる値のこと。

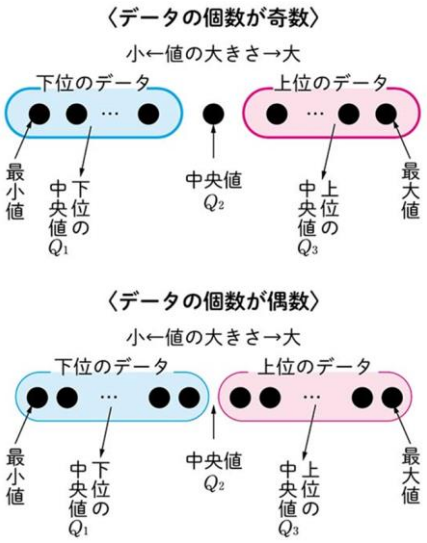
データの個数が偶数のときは、中央に2つの値が並ぶが、その場合は2つの値の平均をとって中央値とする。

(3) 最頻値（モード）

データにおいて、最も個数の多い値のこと。

III. 四分位数

データを値の小さい順に並べたとき、4 等分する位置にくる値を四分位数という。四分位数は、小さい方から第 1 四分位数Q_1、第 2 四分位数Q_2、第 3 四分位数Q_3という。



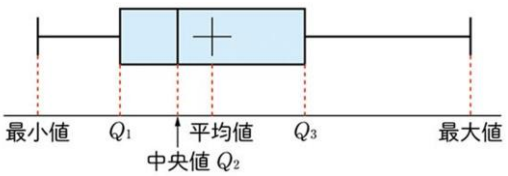
第 1 四分位数Q_1 … 下位のデータの中央値 第 2 四分位数Q_2 … 中央値 第 3 四分位数Q_3 … 上位のデータの中央値

四分位範囲と四分位偏差を以下のように定める。

四分位範囲 $= Q_3 - Q_1$ 四分位偏差 $= \frac{Q_3 - Q_1}{2}$

IV. 箱ひげ図

最小値、最大値、中央値、第 1 四分位数、第 3 四分位数を表したものを箱ひげ図という。作成手順は次のようになる。



① 横軸にデータの値の目盛りをとる。
② Q_1を左端、Q_3を右端とする箱をかき、箱の中に中央値Q_2を示す縦線をかく。
③ 箱の左端から最小値まで、箱の右端から最大値まで線分を引く。
（また、平均値を+の記号で記入することもある）

V. 外れ値

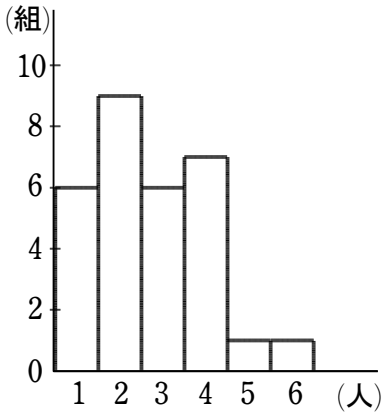
(第 1 四分位数$-1.5\times$四分位範囲)以下の値と、
(第 3 四分位数$+1.5\times$四分位範囲)以上の値のことを外れ値という。

例題 5.1

右のヒストグラムは，ある喫茶店を利用した 30 組について，各組の人数を調べた結果である。

(1)　最頻値，中央値を求めよ。

(2)　平均値を求めよ。



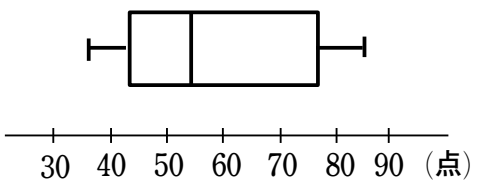
例題 5.2

右の表は，25 人の生徒のテストの得点のデータから作った度数分布表である。このデータの平均値のとり得る範囲を求めよ。

得点の階級(点)	度数
40 以上 49 以下	2
50 ～ 59	5
60 ～ 69	8
70 ～ 79	7
80 ～ 89	3
計	25

例題 5.3

右の図は，ある高校1年生240人に行った期末テストの得点のデータの箱ひげ図である。
この箱ひげ図から読み取れることとして正しいものを，次の①～③から1つ選べ。



① 30点台の生徒は60人である。
② 50点以上の生徒は180人以上いる。
③ 60点未満の生徒は半数以上いる。

例題 5.4

右の表は，10 人の生徒について行った数学と英語のテストの得点のデータを，度数分布表にまとめたものである。また，下の表は，10 人の生徒それぞれについて，テストの得点のデータをまとめたものである。ただし，$a<b$，$c<d$ とする。次の問いに答えよ。

階級 (点) 以上　以下	数学 (人)	英語 (人)
30～39	0	2
40～49	3	3
50～59	4	3
60～69	3	2
合計	10	10

生徒の番号	1	2	3	4	5	6	7	8	9	10	平均値
数学 (点)	41	a	61	57	63	43	b	59	54	50	54
英語 (点)	39	47	35	c	67	d	53	65	55	48	51

(1)　数学の得点のデータの範囲が 25 点であるとき，a，b の値を求めよ。

(2)　英語の得点のデータの中央値を求めよ。

I.　分散と標準偏差

①　変量xのデータ$x_1, x_2, \cdots, x_n$の平均値を$\overline{x}$とするとき、$x_1-\overline{x},\ x_2-\overline{x}, \cdots, x_n-\overline{x}$をそれぞれ、$x_1, x_2, \cdots, x_n$の平均値からの偏差という。また、偏差の2乗の平均値を分散という。さらに、分散の正の平方根を標準偏差といい、sで表す。

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

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

②　変量xの平均値を$\overline{x}$、x^2の平均値を$\overline{x^2}$とするとき、分散s^2と標準偏差sは

$$s^2=\overline{x^2}-(\overline{x})^2 \quad 、\ s=\sqrt{\overline{x^2}-(\overline{x})^2}$$

2乗の平均－平均の2乗

③　$x=au+b$で、uの標準偏差がs'のとき、xの標準偏差sは

$$s=|a|s'$$

II.　仮平均

xのデータ$x_1, x_2, \cdots, x_n$の仮平均をx_0とするとき、xの平均値$\overline{x}$は

$$\overline{x}=x_0+\frac{1}{n}\{(x_1-x_0)+(x_2-x_0)+\cdots+(x_n-x_0)\}$$

III.　共分散

変量x, yのn個の値$x_1, x_2, \cdots, x_n$および、$y_1, y_2, \cdots, y_n$の平均値をそれぞれ$\overline{x}, \overline{y}$とするとき、$x$の偏差と$y$の偏差の積$(x_i-\overline{x})(y_i-\overline{y})$の平均値

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

をxとyの共分散といい、s_{xy}で表す。

IV. 相関係数

xとyの間の関係を図示するために、x, yの値の組を座標とする点を、平面上にとった図を散布図という。

相関関係の強弱をみるために、共分散s_{xy}をx, yの標準偏差s_x, s_yの積で割った量を相関係数といい、rで表す。

$$r = \frac{s_{xy}}{s_x s_y}$$

（rの範囲は$-1 \leqq r \leqq 1$）

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

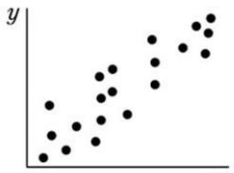
$$= \frac{(x_1-\overline{x})(y_1-\overline{y}) + \cdots + (x_n-\overline{x})(y_n-\overline{y})}{\sqrt{\{(x_1-\overline{x})^2 + \cdots + (x_n-\overline{x})^2\}\{(y_1-\overline{y})^2 + \cdots + (y_n-\overline{y})^2\}}}$$

＜相関係数の性質＞

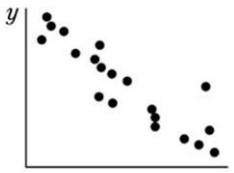
① rの値が1に近いとき、強い正の相関がある。このとき、散布図の点は右上がりの直線にそって分布する傾向が強くなる。

② rの値が-1に近いとき、強い負の相関がある。このとき、散布図の点は右下がりの直線にそって分布する傾向が強くなる。

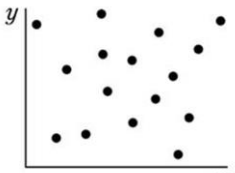
③ rの値が0に近いとき、直線的な相関関係はない(無相関)。



$r \fallingdotseq 1$



$r \fallingdotseq -1$



$r \fallingdotseq 0$

例題 5.5

次のデータは，あるパズルに挑戦した 10 人について，完成するまでにかかった時間 x(分) をまとめたものである。ただし，x のデータの平均値を $\overline{x}$ で表し，20 分を超えた人はいなかったものとする。次の問いに答えよ。

番号	1	2	3	4	5	6	7	8	9	10
x	13	a	7	3	11	18	7	b	16	3
$(x-\overline{x})^2$	4	c	16	64	0	d	16	1	25	64

(1)　$\overline{x}$ の値を求めよ。

(2)　a を b の式で表せ。

(3)　a，b，c，d の値を求めよ。

(4)　x の分散と標準偏差を求めよ。ただし，小数第 1 位を四捨五入せよ。

例題 5.6

20 個の値からなるデータがあり，そのうちの 8 個の値の平均値は 3，分散は 4，残りの 12 個の値の平均値は 8，分散は 9 である。

(1)　このデータの平均値を求めよ。

(2)　このデータの分散を求めよ。

例題 5.7

変量 x の次のデータについて，各問いに答えよ。

672，693，644，665，630，644

(1)　仮平均 x_0 を 630 として，変量 x のデータの平均値 $\overline{x}$ を求めよ。ただし，x のデータの各値と仮平均 x_0 との差の合計を y とすると，求める平均値 $\overline{x}$ は，$\overline{x}=x_0+\dfrac{y}{6}$ で与えられることを用いよ。

(2)　変量 x のデータの分散，標準偏差を求めよ。ただし，正の数 c を用いて，$u=\dfrac{x-x_0}{c}$ とおいて得られる新しい変量 u のデータの標準偏差の c 倍が，変量 x のデータの標準偏差になることを用いよ。

例題 5.8

右の ①，②，③ はある 2 つの変量 x，y のデータについての散布図である。データ ①，②，③ の x と y の相関係数は，0.87，0.04，-0.71 のいずれかである。各データの相関係数を答えよ。