●○ データの分析

2014/6/21 更新

●○●○●○●○●○●○●○●○●○●○●○●○●○

あるクラスで数学の試験を行った結果，女子 10 人の得点は次の通りであった．

72, 65, 76, 51, 67, 72, 80, 54, 56, 47

この例を用いて**【データの分析】**に関する基本用語を解説する．

①のように，実験，調査の結果得られた結果を**データ**という．また，このデータにおける「数学の得点」を**変量**という．以下，この変量を文字 x で表す．また，この**データの大きさ**は 10 である．

この「クラスの女子 10 人」という集団の数学テストにおける傾向・特徴を調べるために ことが**データの分析**の目的である．

A 代表値

(データ①の特徴を表す数値)

データ①を，その特徴がとらえやすいように値の小さい順に並べると次のようになる．

47, 51, 54, 56, 65, 67, 72, 72, 76, 80 $\cdots(*)$

(これらを順に x_1, x_2, x_3, $\cdots x_{10}$ とする．)

これらの総和を個数で割ったものを，変量 x の**平均値**といい，$\overline{x}$ (「エックスバー」と読む) で表す．すなわち

$$\overline{x} = \frac{1}{10}\sum_{k=1}^{10} x_k. \qquad \cdots \frac{\text{変量の値の総和}}{\text{個数}}$$

$$= \frac{x_1 + x_2 + x_3 + \cdots + x_{10}}{10}$$

$$= \frac{47 + 51 + 54 + 56 + 65 + 67 + 72 + 72 + 76 + 80}{10}$$

$$= \frac{640}{10} = 64.$$ (平均値の意味→C〈参考〉を参照)

〈補足〉ここで述べた関係：

$$\text{平均値} = \frac{\text{総和}}{\text{個数}}$$

は，次の形で用いることも多い．

総和 = 平均値 × 個数.

〈参考〉数学 B「数列」を学んでいる人は，上のようにシグマ記号 $\sum$ を使うとデータの分析における様々な量を簡潔な式で表すことができる．ぜひ積極的に使おう．本書では今後，$\sum$ 記号を使わない表現と使う表現を併記する．

5 個　5 個

47, 51, 54, 56, 65, 67, 72, 72, 76, 80

▲

小さい順に並んだデータ $(*)$ において中央の順位 (▲) にくる値を，x の**中央値 (メジアン)** といい，本稿では今後，記号 $\widetilde{x}$ (「エックスチルダ」と読む) で表す．データ $(*)$ ではデータの大きさが偶数なので，▲の位置に値がない．そこで，▲の両隣にある 2 個の値の平均値を中央値とする．すなわち $(*)$ においては

$$\widetilde{x} = \frac{x_5 + x_6}{2} = \frac{65 + 67}{2} = 66.$$

〈注〉上記以外の代表値である「最頻値」については，E「度数分布」で触れる．

B データの散らばり

平均値，中央値だけではわからない「データの散らばり」を調べよう．

データ $(*)$ の**最大値**，**最小値**は，それぞれ

$\max x = 80$, $\min x = 47$.

この 2 つの値の差：

$80 - 47 = 33$

をこのデータの**範囲 (レンジ)** といい，データ全体の分布についての情報が得られる．

データの分布をさらに詳しく表すために次の**四分位数** (Quartile) を考える．

2 個　2 個

47, 51, 54, 56, 65

△　▲

2 個　2 個

67, 72, 72, 76, 80

▲　△

データ $(*)$ を，その中央▲の左側，右側の 5 個ずつのグループに分けたとき，左右各々のグループの中央値 (△) をそれぞれ**第 1 四分位数 (Q_1)**，**第 3 四分位数 (Q_3)** という．$(*)$ では

$Q_1 = 54$, $Q_3 = 72$.

Q_1 は，下から $\frac{1}{4}$ の順位にあたる点数，Q_3 は，上から $\frac{1}{4}$ の順位にあたる点数をそれぞれ表す．また，これら 2 数の差

$Q_3 - Q_1 = 72 - 54 = 18$

を**四分位範囲**といい，全体を成績順に 4 つのグループに分けたとき順位が中央寄りの 2 つのグループにおけるデータの範囲を表す．四分位範囲の $\frac{1}{2}$ を**四分位偏差**という．

また，中央値 $\widetilde{x}$ を**第 2 四分位数 (Q_2)** ともいう．

上で挙げた 5 つの数

47 :最小値 $\min x$

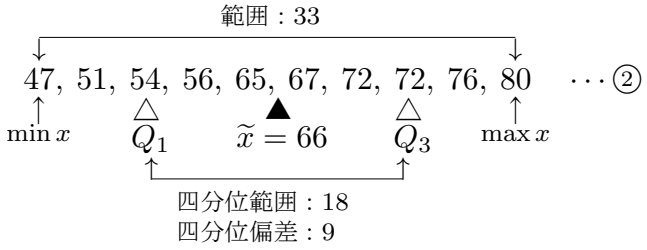
54 :第 1 四分位数 Q_2

66 :中央値 (第 2 四分位数) $\widetilde{x} = Q_2$

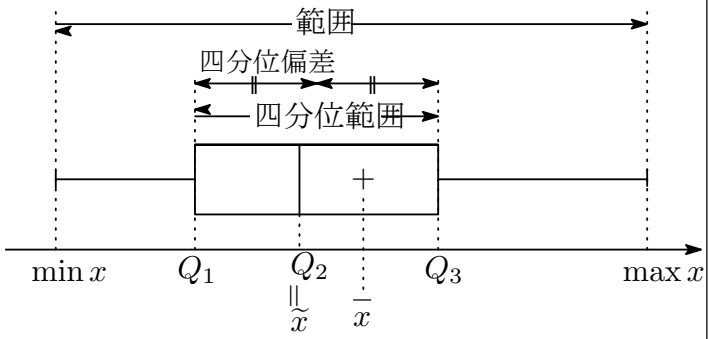
72 :第 3 四分位数 Q_3

80 :最大値 $\max x$

を用いて，データ (∗) の分布・散らばりのおおよその様子をとらえることを**五数要約**という．



五数要約の結果は，次のような**箱ひげ図**で視覚化することができる．



箱ひげ図の“ひげ”: |—— を含めた範囲にデータ全体が，“箱”: □ の部分にそのうち中央値から小・大それぞれの側に $\frac{1}{4}$ ずつ，合わせて全体の半分が分布していることがわかる．

箱ひげ図には，平均値を「＋」の記号で書き入れることもある．

C 偏差

データの散らばりの表し方として，変量の各値の平均値からの“ズレ”に注目する方法もある．

変数 x からその平均値 $\overline{x}$ を引いた値

$$x - \overline{x}$$

を，x の**偏差**という．データ (∗) では次の表のようになる．($\overline{x} = 64$)

x	47	51	54	56	65	67	72	72	76	80
偏差	−17	−13	−10	−8	1	3	8	8	12	16

この偏差に注目してデータの散らばりを表すことを考える．単に偏差を加えるだけでは，正の偏差と負の偏差が打ち消し合ってしまい散らばりを表すことはできない．そこで，偏差を 2 乗したもの (以下,「偏差平方」と呼ぶ) の平均をとった値を考え，これを変量 x の**分散**といい，${s_x}^2$ と表す．すなわち

$$\begin{aligned}
{s_x}^2 &= \frac{1}{10}\sum_{k=1}^{10}\left(x_k - \overline{x}\right)^2 \quad \cdots ①\text{偏差平方の平均}\\
&= \frac{\left(x_1 - \overline{x}\right)^2 + \left(x_2 - \overline{x}\right)^2 + \cdots + \left(x_{10} - \overline{x}\right)^2}{10}\\
&= \frac{1}{10}(289 + 169 + 100 + 64 + 1\\
&\qquad + 9 + 64 + 64 + 144 + 256)\\
&= \frac{1160}{10} = 116.
\end{aligned}$$

分散が大きいほど，データの散らばりが大きいと考えられる．

一般に，n 個からなるデータの分散は，次のように計算することもできる．

$$\begin{aligned}
{s_x}^2 &= \frac{1}{n}\sum_{k=1}^{n}\left(x_k - \overline{x}\right)^2\\
&= \frac{1}{n}\sum_{k=1}^{n}\left({x_k}^2 - 2\overline{x}x_k + \overline{x}^2\right)\\
&= \frac{1}{n}\sum_{k=1}^{n}{x_k}^2 - 2\overline{x}\cdot\frac{1}{n}\sum_{k=1}^{n}x_k + \frac{1}{n}\overline{x}^2\cdot n\\
&= \frac{1}{n}\sum_{k=1}^{n}{x_k}^2 - 2\overline{x}\cdot\overline{x} + \overline{x}^2\\
&= \frac{1}{n}\sum_{k=1}^{n}{x_k}^2 - \overline{x}^2.\\
{s_x}^2 &= \overline{x^2} - \overline{x}^2. \quad \cdots ②(2\text{ 乗の平均})-(\text{平均の 2 乗})
\end{aligned}$$

データ (∗) においては①式の方が簡便だが，②を用いた方がよいこともある．(平均値 $\overline{x}$ が整数値でなく偏差平方の計算が面倒な時とか •••)

分散は，変量の値を 2 乗して得られたものだから，たとえば変量の単位が「cm」ならば，分散の単位は「cm^2」となる．そこで，変量と同じ単位をもつ値として，分散の平方根をとったものを**標準偏差**といい，s_x と表す．データ (∗) では

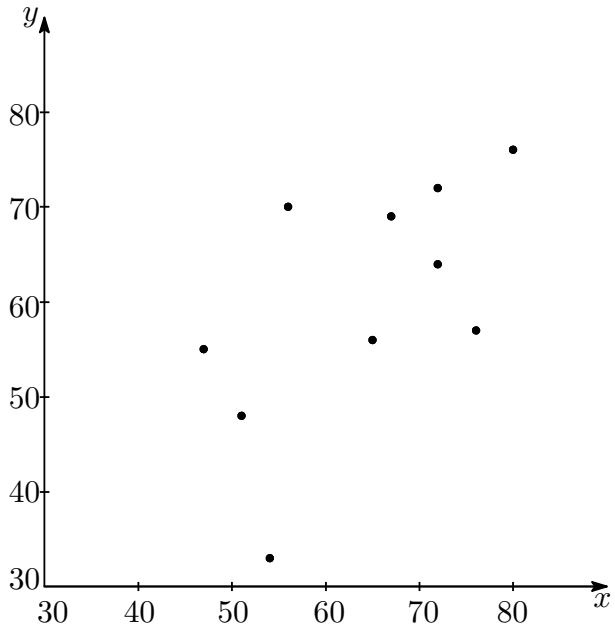
$$s_x = \sqrt{{s_x}^2} = \sqrt{116} = 34.05\cdots.$$

D データの相関

数学の得点データ (∗) を得たクラスで，英語の得点も調べたところ，次の表のようなデータ (∗∗) を得た．(数学の得点が低い順に並べてある．)

k	1	2	3	4	5	6	7	8	9	10
x_k	47	51	54	56	65	67	72	72	76	80
y_k	55	48	33	70	56	69	64	72	57	76

データ (∗∗) における 2 つの変量の関係を直観的に把握するには，座標平面上に各生徒に対応する点 $(x,\ y)$ をとって得られる**散布図** (相関図) を描くとよい．(実際にはもっとデータの大きさが大きいときこそ有効なのだが)



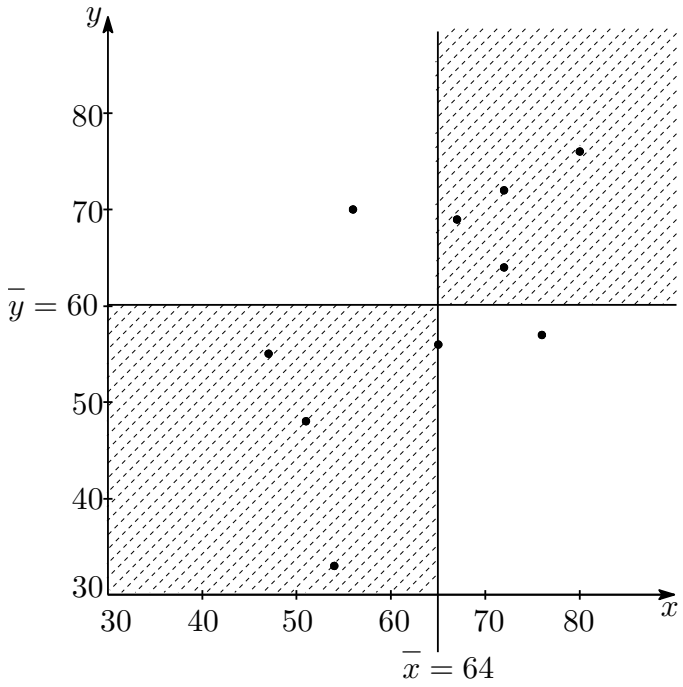
x と y の間の関係の強さを数量的にとらえるには，それぞれの偏差どうしの関係に注目する．

k	x_k	y_k	$x_k-\overline{x}$	$y_k-\overline{y}$	偏差積
1	47	55	−17	−5	85
2	51	48	−13	−12	156
3	54	33	−10	−27	270
4	56	64	−8	10	−80
5	65	56	1	−4	−4
6	67	69	3	9	27
7	72	64	8	4	32
8	72	72	8	12	96
9	76	57	12	−3	−36
10	80	76	16	16	256
	平均値 $\overline{x}=64$	平均値 $\overline{y}=60$	標準偏差 $s_x=10.8$	標準偏差 $s_y=12.3$	共分散 $s_{xy}=80.2$

この表中にある偏差積 : $\left(x_k-\overline{x}\right)\left(y_k-\overline{y}\right)$ が正・負のどちらに偏っているかにより，$x,\ y$ の偏差 (平均値からの“ズレ”) が同傾向・逆傾向のいずれであるかが判断できる．

偏差積が正 $\leftrightarrow$ $x,\ y$ の偏差が同符号 $\leftrightarrow$ $x,\ y$ が同傾向

偏差積が負 $\leftrightarrow$ $x,\ y$ の偏差が異符号 $\leftrightarrow$ $x,\ y$ が逆傾向



このデータ (∗∗) では，上記散布図でグレー部分にある生徒の偏差積は正，白色部分にある生徒の偏差積は負であり，偏差積はかなり正に偏っている．

$x,\ y$ の偏差どうしの積 $\left(x-\overline{x}\right)\left(y-\overline{y}\right)$ の平均値を**共分散**といい，s_{xy} で表す．すなわち

$$\begin{aligned} s_{xy} &= \frac{1}{n}\sum_{k=1}^{n}\left(x_k-\overline{x}\right)\left(y_k-\overline{y}\right) \\ &= \frac{1}{n}\Big\{\left(x_1-\overline{x}\right)\left(y_1-\overline{y}\right)+\left(x_2-\overline{x}\right)\left(y_2-\overline{y}\right)+ \\ &\qquad \cdots+\left(x_n-\overline{x}\right)\left(y_n-\overline{y}\right)\Big\}. \end{aligned}$$

この共分散が正・負のどちらに偏っているかにより，$x,\ y$ の変化が同傾向か逆傾向かがわかる．ただし，共分散の値は変数の単位を変えるなどしただけで変化してしまうので，同じように変化する標準偏差で割ることにより，$x,\ y$ の関係の強さを普遍的に表す数値を作る．これが，x と y の**相関係数** r_{xy} である．すなわち

$$r_{xy}=\frac{s_{xy}}{s_x s_y}.$$

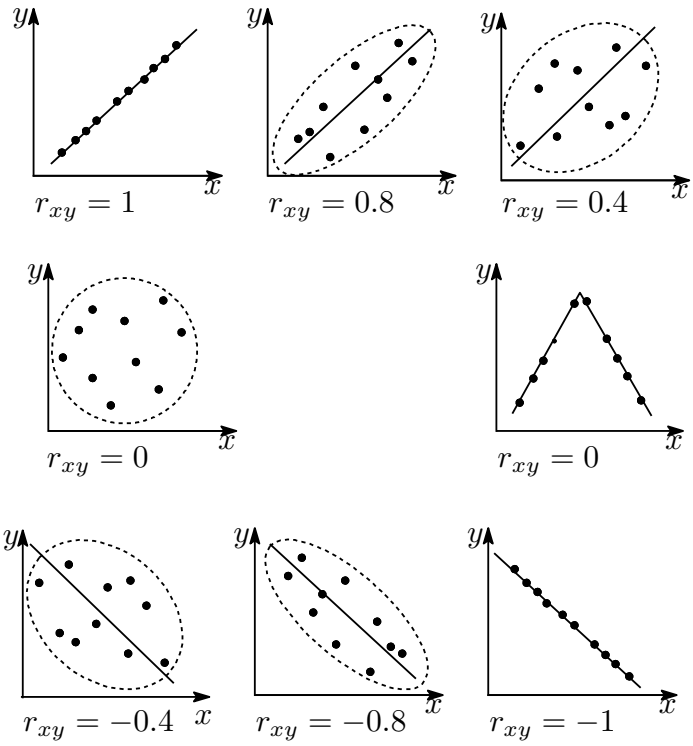
データ (∗∗) においては，表中の数値を用いて

$$r_{xy}=\frac{s_{xy}}{s_x s_y}=\frac{80.2}{10.8\times 12.3}=0.60$$

である．

相関係数の値は，$-1\leq r_{xy}\leq 1$ の範囲にある．r_{xy} が $+1$ に近いときは，偏差どうしの積の値が正に偏っている．つまり，x の増加にともない y も増加する傾向が強い．このとき $x,\ y$ の間には**正の相関**があるという．逆に r_{xy} が -1 に近いとき，x の増加にともない y は減少する傾向が強く，**負の相**

関があるという．また，r_{xy} が 0 に近いときは「相関がない」という．



〈注意 1〉 相関係数は，散布図における各点が右上がりのある直線に近くに分布しているほど 1 に近い値をとり，右下がりの直線近くに分布しているほど -1 に近くなる．相関係数は，あくまでも直線の近くに分布している度合いを表す数値であり，たとえば上図中段右において相関係数は 0 となるのだが，x, y の間に何の関係もないとは言い難い．

〈注意 2〉 「相関関係」は「因果関係」とは異なる．たとえば x, y の間に正の相関がある，つまり x の増加にともない y も増加する傾向があるとしても，x，y のいずれかが他方の「原因」であるとは言い切れない．これは，x, y が共通な要因 z によって影響を受け，結果として x, y に正の相関が生まれた可能性もあるからである．

〈参考〉 相関係数が $-1 \leqq r_{xy} \leqq 1$ の範囲にあることは，次のようにして示される．(以下，$\sum_{k=1}^{n}$ を単に $\sum$ と記す．

$a_k = x_k - \overline{x}$, $b_k = y_k - \overline{y}$ とおき，さらに $A = \sqrt{\sum {a_k}^2}$, $B = \sqrt{\sum {b_k}^2}$ とおくと，

$$r_{xy} = \frac{\frac{1}{n}\sum a_k b_k}{\sqrt{\frac{1}{n}\sum {a_k}^2}\sqrt{\frac{1}{n}\sum {b_k}^2}} = \frac{\sum a_k b_k}{AB} \cdots ㋐$$

$\sum \left(\frac{a_k}{A} \pm \frac{b_k}{B}\right)^2 \geqq 0$ より，

$$\frac{\sum {a_k}^2}{A^2} \pm 2\frac{\sum a_k b_k}{AB} + \frac{\sum {b_k}^2}{B^2} \geqq 0.$$

$2 \pm 2r_{xy} \geqq 0$.　$\therefore\ -1 \leqq r_{xy} \leqq 1$.□

ちなみに，この不等式を変形すると次のようになる．

$$-AB \leqq \sum a_k b_k \leqq AB.$$

$$\left(\sum a_k b_k\right)^2 \leqq \left(\sum {a_k}^2\right)\left(\sum {b_k}^2\right).$$

これは，「コーシーシュワルツの不等式」と呼ばれる有名不等式の一般形である．

〈参考 1〉 相関係数の意味

上記㋐式で，$n = 3$ のときを考える．x, y の偏差を成分とする 2 ベクトル $\vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$, $\vec{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$ を用いて㋐を表すと

$$r_{xy} = \frac{a_1 b_1 + a_2 b_2 + a_3 b_3}{\sqrt{{a_1}^2 + {a_2}^2 + {a_3}^2}\sqrt{{b_1}^2 + {b_2}^2 + {b_3}^2}}$$
$$= \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|}$$
$$= \cos\theta. \ (\theta \text{は 2 ベクトルのなす角})$$

つまり，相関係数とは，偏差を成分とするベクトルどうしのなす角の余弦であり，次の関係がある．

$r_{xy} = \cos\theta$ が 1 に近いほど，θ が 0 に近い，
すなわち x, y の偏差が同傾向
$r_{xy} = \cos\theta$ が -1 に近いほど，θ が π に近い，
すなわち x, y の偏差が逆傾向

E 変量の変換

(以下，a, b は定数とする．)

(1)　変量 x と，それから作られる別の変量の平均値，分散，標準偏差の関係は次の通り．

	平均値	分散	標準偏差	
x	$\overline{x}$	${s_x}^2$	s_x	
ax	$a\overline{x}$	$a^2{s_x}^2$	$\lvert a\rvert s_x$	$\cdots$①
$x+b$	$\overline{x}+b$	${s_x}^2$	s_x	$\cdots$②
$ax+b$	$a\overline{x}+b$	$a^2{s_x}^2$	$\lvert a\rvert s_x$	$\cdots$③

〈注意 2〉 共分散 s_{xy} は，x, y の一方に上記①～③の変換を施した際，a が正なら標準偏差 s_x と同じ変化をし，相関係数 r_{xy} は変化しない (a が負なら符号が反対になる)．

〈参考〉③を証明する．(ここでも $\sum_{k=1}^{n}$ を $\sum$ と記す．)

$$\begin{aligned}\overline{ax+b} &= \frac{1}{n}\sum(ax_k+b)\\ &= \frac{1}{n}\left(a\sum x_k + bn\right)\\ &= a\cdot\frac{1}{n}\sum x_k + b = a\overline{x}+b.\end{aligned}$$

$$\begin{aligned}{s_{ax+b}}^2 &= \frac{1}{n}\sum\left\{(ax_k+b)-\left(\overline{ax+b}\right)\right\}^2\\ &= \frac{1}{n}\sum\left\{(ax_k+b)-\left(a\overline{x}+b\right)\right\}^2\\ &= \frac{1}{n}\sum\left\{a\left(x_k-\overline{x}\right)\right\}^2\\ &= a^2\frac{1}{n}\sum\left(x_k-\overline{x}\right)^2 = a^2{s_x}^2.\end{aligned}$$

$$s_{ax+b} = \sqrt{a^2{s_x}^2} = |a|s_x.\square$$

(①，②についても同様に示せる．)

(2)　変量 x の平均値 $\overline{x}$ を求める際，平均値に近そうで計算しやすい値 X を選び，$x-X$ の平均値を用いて $\overline{x}$ を求めると簡便である．

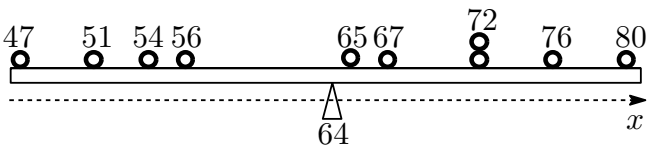
$$\begin{aligned}\overline{x} &= \overline{(x-X)+X}\\ &= \overline{x-X}+X \qquad \cdots \text{前記②を利用}\\ &= \frac{1}{n}\sum_{k=1}^{n}(x-X)+X\\ &= \frac{(x_1-X)+(x_2-X)+\cdots+(x_n-X)}{n}+X.\end{aligned}$$

このとき X を**仮平均**という．

〈参考〉同様にして，$x-\overline{x}$ の平均は

$$\overline{x-\overline{x}} = \overline{x}-\overline{x} = 0.$$

〈参考〉偏差の平均値は必ず 0 となる (証明は，後述する E ②式による)．つまり偏差の総和は常に 0 である．このことを用いて，「平均値」の物理的・視覚的意味を説明する．



質量の無視できる板を数直線に沿って置き，x の平均値に対応する位置に支点をとる．また，データの各値 x 毎に，対応する位置に質量 1 の重りを乗せる．このとき，右回りを正とする「力のモーメント」の総和は

$$\sum_{k=1}^{10}\left(x-\overline{x}\right)\cdot 1 = 0.$$

この等式は，偏差が正で支点より右にある重りによる右回りのモーメントと，偏差が負で支点より左にある重りによる左回りのモーメントとが釣り合っていること，すなわち，これら 10 個の重りからなる物体の重心と平均値が一致することを表している．

(3)　変量 x から，その平均値と標準偏差を用いて作られる変量 $z = \dfrac{x-\overline{x}}{s_x}$ は，上記①〜③より $\overline{z}=0$, $s_z=1$ を満たす．このような変量の変換を**基準化**もしくは**標準化**という．

この基準化された z を用いて任意の平均値・標準偏差をもつ変量を作ることができる．たとえば $Z = 10z+50$ は，③より $\overline{Z}=50$, $s_Z=10$ を満たす変量である．この Z が，受験生お馴染みの**偏差値**である．

F 度数分布

(1)　データの大きさ n が大きいとき，変量の値を同じ幅 (**階級の幅**) のいくつかの区間 (**階級**) に分け，各階級に対してその区間に入っているものの個数 (**度数**) を対応させた**度数分布**を考えるとよい．これを表にしたものを**度数分布表**という．また，各階級値の n に対する割合：$\dfrac{\text{度数}}{n}$ を**相対度数**という．

各階級には，実際には異なる値が含まれているが，その違いを無視し，全ての値が階級の真ん中の値 (**階級値**) をとるものとみなせば様々な計算を簡略化できる．

(例) あるクラスの生徒の通学時間 (単位：分)

階級の幅：20，データの大きさ：50

階級 以上〜未満	階級値	度数	相対度数
0〜20	10	3	0.06
20〜40	30	14	0.28
40〜60	50	23	0.46
60〜80	70	8	0.16
80〜100	90	2	0.04

(2)　度数を棒グラフで表したものを**ヒストグラム**という．

(3)　度数が最大である階級の階級値を**最頻値** (**モード**) という．上の例では，最頻値は 50 である．