3. 1次変換

1. 1次変換

変数 x から新しい変数 y をつくることを，x を y に変数変換するという。統計学では，1 次変換という変数変換が重要になる。(線形代数における「1次変換」とは異なる。)

一般に，変数 x を x の 1 次式

$$y = ax + b \quad (a, b \text{は定数})$$

によって，y に変数変換することを，「x の 1 次変換」または「x を 1 次変換する」という。

データの場合，与えられた n 個のデータ $x_1, x_2, \cdots, x_n$ に対して，新しい n 個のデータ

$$y_1 = ax_1 + b, \quad y_2 = ax_2 + b, \quad \cdots, \quad y_n = ax_n + b$$

をつくることを，データ x_i の 1 次変換という。

新データ y_i による計算は，xy 平面において，直線 $y = ax + b$ 上の点 (x_i, y_i) を考え，この点の y 座標 y_i で計算することを意味する。

次の 1 次変換の公式は重要である。

● 1次変換の公式

n 個のデータ $x_1, x_2, \cdots, x_n$ に対して，a, b を定数として

$$y_i = ax_i + b \quad (i = 1, 2, \cdots, n)$$

とおくと，$y_1, y_2, \cdots, y_n$ について，以下が成り立つ。

(1) $\bar{y} = a\bar{x} + b$ $\quad$ y_i の平均 $=$ a $\times$ (x_i の平均) $+$ b

(2) $y_i - \bar{y} = a(x_i - \bar{x})$ $\quad$ y_i の偏差 $=$ a $\times$ (x_i の偏差)

(3) $\sigma_y^{\,2} = a^2\sigma_x^{\,2}$ $\quad$ y_i の分散 $=$ a^2 $\times$ (x_i の分散)

(4) $\sigma_y = |a|\sigma_x$ $\quad$ y_i の標準偏差 $=$ $|a|$ $\times$ (x_i の標準偏差)

特に $a > 0$ のときは，$\sigma_y = a\sigma_x$

(証明)

(1) $\bar{y}$ は $y_1, y_2, \cdots, y_n$ の平均なので，

$$\begin{aligned}
\bar{y} &= \frac{1}{n}(y_1 + y_2 + \cdots + y_n) \\
&= \frac{1}{n}\left\{(ax_1 + b) + (ax_2 + b) + \cdots + (ax_n + b)\right\} \\
&= \frac{1}{n}\left\{(ax_1 + ax_2 + \cdots + ax_n) + (b + b + \cdots + b)\right\} \\
&= \frac{1}{n}\left\{a(x_1 + x_2 + \cdots + x_n) + nb\right\}
\end{aligned}$$

$$= a\cdot\frac{1}{n}(x_1 + x_2 + \cdots + x_n) + \frac{1}{n}\cdot nb$$
$$= a\cdot\bar{x} + b$$

(2) (1)より,

$$y_i - \bar{y} = (ax_i + b) - (a\bar{x} + b) = a(x_i - \bar{x})$$

(3) (2)より, $(y_i - \bar{y})^2 = a^2(x_i - \bar{x})^2$ なので,

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

(4) (3)より,

$$\sigma_y{}^2 = \sqrt{a^2\cdot\sigma_y{}^2} = |a|\sigma_y$$

(証明終)

(4)では, a の値は正とは限らないので, 絶対値 $|a|$ になる。従って

$a > 0$ のときは $\sigma_y = a\sigma_x$

分散や標準偏差は, 各データの偏差で決まる量である。$a = 1$ の場合は, (2)により, 1 次変換を行っても各データの偏差は変化しないので, 分散や標準偏差の値も変わらないことに注意しよう。

なお, 1 次変換の場合, 変換後のデータの度数分布曲線は, 変化はするが元の曲線の形が縦軸方向に拡大または縮小するように変化する。そのため, 元のデータが正規分布をなしていれば, 1 次変換を行っても正規分布をなす。

2. 標準化変換

標準化変換は 1 次変換であり, どのようなデータも標準化変換を行えば, 平均を 0, 標準偏差を 1 にすることができる。

● 標準化変換

n 個のデータ $x_1, x_2, \cdots, x_n$ に対して

$$z_i = \frac{x_i - \bar{x}}{\sigma_x} = \frac{x_i \text{の偏差}}{x_i \text{の標準偏差}} \quad (i = 1, 2, \cdots, n)$$

とおく。この 1 次変換を標準化変換といい, z_i を x_i の標準測度(標準化変量)という。また, x_i を z_i に変換することを「x_i を標準化する」という。

x_i が得点のデータのときは，標準測度を，「標準得点」「$z-$評点」「$z-$score」などという。

標準測度 $z_1, z_2, \cdots, z_n$ について，以下が常に成り立つ。

(1) 平均 $\bar{z} = 0$ （標準測度の平均は 0）

(2) 分散 $\sigma_z^{\,2} = 1$，標準偏差 $\sigma_z = 1$ （標準測度の分散や標準偏差は 1）

(3) 合計 $\sum_{i=1}^{n} z_i = z_1 + z_2 + \cdots + z_n = 0$ （標準測度の合計は 0）

(4) 平方和 $\sum_{i=1}^{n} z_i^{\,2} = z_1^{\,2} + z_2^{\,2} + \cdots + z_n^{\,2} = n$ （標準測度の平方和は n ）

上記の証明は容易である。変換の式は

$$z_i = \frac{x_i - \bar{x}}{\sigma_x} = \left(\frac{1}{\sigma_x}\right) x_i + \left(-\frac{\bar{x}}{\sigma_x}\right)$$

と変形でき，これは $z_i = a x_i + b$ （a, b は定数）の形をしているので，1 次変換である。

よって，1 次変換の公式を用いれば

$$\bar{z} = a\bar{x} + b = \left(\frac{1}{\sigma_x}\right)\bar{x} + \left(-\frac{\bar{x}}{\sigma_x}\right) = 0$$

$$\sigma_z^{\,2} = a^2 \sigma_x^{\,2} = \left(\frac{1}{\sigma_x}\right)^2 \sigma_x^{\,2} = 1$$

となる。従って，

$$\sigma_z = \sqrt{\sigma_z^{\,2}} = \sqrt{1} = 1, \qquad \sum_{i=1}^{n} z_i = n \times \bar{z} = 0$$

また，分散 $\sigma_z^{\,2}$ は 1 であるから，分散の定義式に戻れば

$$1 = \sigma_z^{\,2} = \frac{1}{n}\sum_{i=1}^{n} (z_i - \bar{z})^2 = \frac{1}{n}\sum_{i=1}^{n} z_i^{\,2} \qquad \therefore \sum_{i=1}^{n} z_i^{\,2} = n$$

上記では，どのようなデータも，標準化すれば，標準測度の平均は 0，標準偏差は 1 になることを示している。一方，標準化変換は 1 次変換であるから，元のデータが正規分布をなしていれば，標準測度のデータも正規分布をなし，この正規分布の平均は 0，標準偏差は 1 である。

あとで学習するが，正規分布といってもいろいろな形があり，山の高い正規分布，山の低い正規分布がある。また，同じ形の正規分布でも，平均が異なれば，分布の位置も異なる。

しかし，正規分布の形と位置は，平均と標準偏差の 2 つの値で完全に決まり，特に，平均が 0，標準偏差が 1 の正規分布を「標準正規分布」という。従って，正規分布をなしているデータは，どんなデータであっても，それを標準化すれば，その標準測度のデータは，形と位置が

確定した標準正規分布になる。

この事実は重要であり，応用範囲も広い。例えば，英語の点数と数学の点数がともに正規分布をなすとき，両者の正規分布は通常，形も位置も異なるので，英語の 65 点と数学の 60 点のどちらの成績が良いかは，点数だけでは判断できない。しかし，両者のデータをそれぞれ標準化すれば，どちらも標準正規分布になり，分布が完全に一致する。よって，英語 65 点の標準得点と，数学 60 点の標準得点を比較して，標準得点の大きい方が成績が良いと判断できる。

3. 名数と無名数

復習である。名数（めいすう）と無名数（むめいすう）は，小学校で次のように学習する。

① 数に量の単位名をつけて表した数を「名数」という。それに対して，単位名のついてない数を「無名数」という。

○ 名数の例 ：1 枚，2 羽，3 冊，4 個，5 台，6m，7kg，8cm²，9L，10 時間

○ 無名数の例：1，2，3，4，5，6，7，8，9，10

② 名数と無名数の乗除の計算は，次のようになる。

名数 × 無名数 ＝ 名数　　名数 ÷ 無名数 ＝ 名数　　名数 ÷ 名数 ＝ 無名数

○ 例： 60cm × 3 ＝ 180cm，　8kg ÷ 4 ＝ 2kg，　16m² ÷ 2m² ＝ 8

標準測度（標準得点）は，無名数である。例えば，データ x_i の単位が cm であれば，標準偏差 σ_x の単位も cm であるから，標準測度 z_i の単位は cm／cm となり， z_i には単位がない。

4. 正規分布での割合

正規分布は平均を中心にして左右対称の山型の分布をなし，数学的に定義された曲線であるので，区間に入るデータの割合は正確に計算できる。

データ x_i の分布が正規分布をなすとき，

$\bar{x} - \sigma_x \sim \bar{x} + \sigma_x$　すなわち　平均 － 標準偏差 ～ 平均 ＋ 標準偏差

の範囲にあるデータの割合は約 68.3%であるが，大ざっぱに 70%と理解しておけばよい。以下の割合（70%，95%，100%）は，覚えておくと良い。

● 正規分布の場合

① $\bar{x} - \sigma_x \sim \bar{x} + \sigma_x$ の範囲にあるデータの割合は約 70%

② $\bar{x} - 2\sigma_x \sim \bar{x} + 2\sigma_x$ の範囲にあるデータの割合は約 95%

③ $\bar{x} - 3\sigma_x \sim \bar{x} + 3\sigma_x$ の範囲にあるデータの割合は約 100%

一般に，身長，体重，試験の点数などは，データ数が多いほど，正規分布に近づくことが経験的に知られている。

■ **例1**

平均が 60 点，標準偏差が 10 点の試験において，太郎の成績 65 点の標準得点を求めよ。

（解） $\dfrac{65-60}{10} = 0.5$

■ **例 2**

200 人の生徒に対して，英語と数学の試験を実施したところ，以下のような結果になった。生徒の太郎の成績は，英語が 65 点，数学が 60 点であった。

	平均点	標準偏差	太郎の成績
英語	50	16	65
数学	40	20	60

(1) この表の結果だけから判断した場合，太郎の英語の成績と数学の成績は，成績の順位としてはどちらが良いといえるか。

(2) また，太郎の数学の成績 60 点は，上位から何番目といえるか。

＜考え方＞

ここでは，以下のように大ざっぱに考えてよい。（詳細な議論は後回し）

① データは試験の点数であるから，英語の成績（200 個の英語の点数）の分布は，だいたい山型になると考えられる。数学の成績（200 個の数学の点数）も同様である。

② データ数は 200 であるので，かなり多い。従って，英語の成績の分布は，正規分布に近いものになると考えてよい。数学の成績についても同様である。（一般に，データ数が多いほど，正規分布に近づく。）

③ そこで，英語の成績は，正規分布をなすと仮定する。数学の成績も同様である

④ このとき，英語の成績の標準得点（200 個の英語の点数の標準得点）は，標準正規分布をなす。数学の成績の標準得点（200 個の数学の点数の標準得点）も同様である。その結果，

英語の成績の標準得点の分布 ＝ 標準正規分布 ＝ 数学の成績の標準得点の分布

⑤ これにより，太郎の英語の成績と数学の成績については，標準得点の大きい方が，成績の順位が良いと判断できる。

(1) 太郎の成績について

英語 65 点の標準得点は，$\dfrac{65-50}{16} \fallingdotseq 0.94$

数学 60 点の標準得点は，$\dfrac{60-40}{20} = 1.00$

$0.94 < 1.00$ であるから，成績順位としては数学の成績の方が良いと考えられる。

(2) 数学の成績の分布は正規分布と仮定しているので,

$$40 - 20 \sim 40 + 20 \quad \text{すなわち } 20\text{点}\sim 60\text{点}$$

の範囲にあるデータの割合（数学の点数の割合）は約 70%である。よって，60 点以上のデータの割合は 15%になるから，数学の点数が 60 点以上の生徒の人数は

$$200 \times 0.15 = 30 \text{（人）}$$

従って，太郎の数学の成績 60 点は，上位から 30 番目といえる。

5. 偏差値

データ x_i が試験の点数の場合，その標準得点 z_i の平均は 0 である。しかし，「平均が 0」では考えにくいこともあるので，標準得点を 10 倍して，50 を加えた値を考える。これが偏差値である。

● 偏差値の定義

n 個の点数 $x_1, x_2, \cdots, x_n$ に対して

$$T_i = \frac{x_i - \bar{x}}{\sigma_x} \times 10 + 50 = (x_i \text{の標準得点}) \times 10 + 50 \quad (i = 1, 2, \cdots, n)$$

とおくと，$T_1, T_2, \cdots, T_n$ について常に次が成り立つ。

$$\text{平均 } \overline{T} = 50, \quad \text{標準偏差 } \sigma_T = 10$$

T_i を x_i の偏差値という。

上記では，x_i の標準得点 z_i を，さらに 1 次変換

$$T_i = 10z_i + 50 \qquad \left(z_i = \frac{x_i - \bar{x}}{\sigma_x}\right)$$

によって T_i に変数変換している。よって，$\bar{z} = 0$，$\sigma_z = 1$ より

$$\overline{T} = 10\bar{z} + 50 = 50, \quad \sigma_T = 10\sigma_z = 10$$

偏差値の平均は必ず 50 であり，標準偏差は 10 である。元の点数 x_i が正規分布をなせば，標準得点は平均 0，標準偏差 1 の標準正規分布をなし，偏差値は平均 50，標準偏差 10 の正規分布をなす。成績を比較するときは，標準得点のかわりに，偏差値で比較してもよい。

なお，偏差値も無名数であるので，「偏差値が 55 点」などという言い方をしてはいけない。

■ 例

平均が 53 点，標準偏差が 12 点の試験において，太郎の成績 65 点の偏差値を求めよ。

（解） $\dfrac{65 - 53}{12} \times 10 + 50 = 60$

＜注意＞

1 次変換を何度行っても，その全体は 1 つの 1 次変換である。例えば，

$$y_i = ax_i + b \quad (a, b \text{は定数})$$
$$w_i = cy_i + d \quad (c, d \text{は定数})$$

とおけば，

$$w_i = c(ax_i + b) + d = cax_i + cb + d$$

つまり，$x_i \rightarrow y_i \rightarrow w_i$ のように 1 次変換を繰り返しても，最後の w_i は最初の x_i を 1 次変換したものである。従って，試験の点数の偏差値は，点数を 1 次変換したものである。

6. 変動係数

標準偏差や分散は，データのバラツキの度合いを表す量である。しかし，2 種類のデータのバラツキを比較するときに，単位が異なる場合や，平均がかなり違う場合は，単純に標準偏差や分散の値で比較することはできない。

例えば，右のような身長と体重のデータがあるとき，標準偏差の数値を比較して

$$6.5 > 3.0$$

であるから，身長の方が体重よりもバラツキが大きいとは言えない。体重の単位を g にすれば，3.0kg は 3000g になり，$6.5 < 3000$ になってしまう。

	平均	標準偏差
身長	160 cm	6.5 cm
体重	50 kg	3.0 kg

また，右のデータにおいて，単位は同じだが，

$$250000 > 3$$

であるから，象の体重の方がバラツキが大きいなどとは言えない。平均がかなり違う。

	平均	標準偏差
象の体重	5000000 g	250000 g
ねずみの体重	30 g	3 g

このような場合，次の変動係数でバラツキを比較できる。

● 変動係数（変異係数）

データの標準偏差 σ_x を平均 $\bar{x}$ で割った値を，データの「変動係数」または「変異係数」といい，CV で表す。すなわち，

$$CV = \frac{\sigma_x}{\bar{x}} = \frac{\text{標準偏差}}{\text{平均}}$$

2 種類のデータを比較するとき，変動係数の大きい方がバラツキが大きいといえる。

データ $x_1, x_2, \cdots, x_n$ の平均 $\bar{x}$ の値が正のとき，$y_i = \frac{1}{\bar{x}} \cdot x_i$ という 1 次変換を行えば，

$$\bar{y} = \frac{1}{\bar{x}} \cdot \bar{x} = 1, \quad \sigma_y = \frac{1}{\bar{x}} \cdot \sigma_x = CV$$

となる。つまり，各データを平均＞0 で割れば必ず平均は 1 になり，そのときの標準偏差が変動係数である。定義から，変動係数も無名数である。

変動係数は，平均を 1 にした場合のバラツキを表す量である。2 種類のデータを比較するとき，平均がかなり異なる場合，あるいは単位が異なる場合は，上記の 1 次変換によって両者の平均をともに 1 にして一致させれば，そのときの両者の標準偏差（＝変動係数）でバラツキが比較できる。そして，変動係数が大きいほど，データのバラツキが大きいといえる。

最初の例では

$$\text{身長の変動係数}\quad CV = 6.5/160 \fallingdotseq 0.04$$
$$\text{体重の変動係数}\quad CV = 3.0/50 = 0.06$$

であるから，体重の方がバラツキが大きいといえる。