3　分散と標準偏差

3.1　分散と標準偏差

分布の広がりを表す数値のことを**散布度**といい，分散，標準偏差，範囲，四分位偏差等がある。次のデータは男性 25 人の体重である。

59	64	58	68	51	63	57	66	55	54	64	56	74	64	68	
61	56	59	64	59	73	65	57	61	69						

偏差　観測値 $x_1, x_2, \cdots, x_n$ の平均値を $\bar{x}$ とし，各観測値 x_i から平均値 $\bar{x}$ を引いた値を偏差（平均値からの偏差）という。

$$x_1 - \bar{x}, \quad x_2 - \bar{x}, \quad \cdots, \quad x_n - \bar{x}$$

分散（variance）　偏差の平方の平均値のことを分散といい，s_x^2 あるいは簡単に s^2 と表す。[*i]

[*i]

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}, \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (1)$$

上のデータの分散を定義式 (1) で求める。平均値は $\bar{x} = 1545/25 = 61.8$ である。

$$s^2 = \frac{(59 - 61.8)^2 + \cdots + (69 - 61.8)^2}{25} = 33.12$$

分散（2）　分散については公式 $s^2 = \overline{x^2} - \bar{x}^2$ がよく知られている。

$$\begin{aligned} s^2 &= \frac{1}{n}\sum (x_i - \bar{x})^2 = \frac{1}{n}\sum ({x_i}^2 - 2\bar{x}x_i + \bar{x}^2) \\ &= \frac{1}{n}\sum {x_i}^2 - 2\bar{x}\frac{1}{n}\sum x_i + \frac{1}{n}\sum \bar{x}^2 \\ &= \overline{x^2} - 2\bar{x}\cdot\bar{x} + \bar{x}^2 \\ &= \overline{x^2} - \bar{x}^2 \end{aligned}$$

定義式 (1) のかわりに，(2) を用いてもよい。

$$s^2 = \frac{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2}{n} - \bar{x}^2, \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}{x_i}^2 - \bar{x}^2 \qquad (2)$$

上のデータの分散を公式 (2) で求める。

$$s^2 = \frac{59^2 + 64^2 + \cdots + 69^2}{25} - 61.8^2 = 33.12$$

[*i] 本稿では，n で割るものを分散とよぶ。

分散（3）　公式 (2) をさらに変形した (3) を用いることもできる。

$$s^2 = \frac{n\left(x_1^2 + \cdots + x_n^2\right) - \left(x_1 + \cdots + x_n\right)^2}{n^2}, \quad s^2 = \frac{n\sum {x_i}^2 - \left(\sum x_i\right)^2}{n^2} \tag{3}$$

上のデータの分散を公式 (3) で求める。

$$\begin{aligned} s^2 &= \frac{25\left(59^2 + \cdots + 69^2\right) - \left(59 + \cdots + 69\right)^2}{25^2} \\ &= \frac{25 \cdot 96309 - 1545^2}{25^2} = 33.12 \end{aligned}$$

標準偏差（SD; standard deviation）　分散の単位はもとの観測値の単位の 2 乗になっているが，それの平方根を求めると，もとの単位に戻すことができる。分散 s_x^2 の（正の）平方根のことを標準偏差といい，s_x あるいは簡単に s と表す。

$$s = \sqrt{s^2}, \qquad \text{標準偏差} = \sqrt{\text{分散}}$$

上のデータの標準偏差は $s = \sqrt{33.12} = 5.75$ である。

変動係数（CV; coefficient of variation）　平均値が異なる 2 つの集団に対して，標準偏差を比較するのは難しい場合がある。標準偏差を平均値で割った値のことを変動係数または変異係数という。変動係数は比率尺度のデータで使用する。

$$\mathrm{CV} = \frac{s_x}{\bar{x}}$$

3.2　度数分布表から分散を求める

次の表は男性 25 人の体重から作った分布表である。

階級	階級値 x	度数 f	相対度数 p
$47.5 \sim 52.5$	50	1	0.04
$52.5 \sim 57.5$	55	6	0.24
$57.5 \sim 62.5$	60	6	0.24
$62.5 \sim 67.5$	65	7	0.28
$67.5 \sim 72.5$	70	3	0.12
$72.5 \sim 77.5$	75	2	0.08

平均値　度数 f_j の総和を n とする。平均値を求めるとき，度数 f_j を用いる方法，相対度数 p_j を用いる方法がある。

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} x_j f_j, \qquad \bar{x} = \sum_{j=1}^{k} x_j p_j$$

上のデータの平均値を度数分布表から求める。

$$\bar{x} = \frac{50 \cdot 1 + 55 \cdot 6 + \cdots + 75 \cdot 2}{25} = \frac{1555}{25} = 62.2$$

分散（1）　分散を求めるとき，度数を用いる方法，相対度数を用いる方法がある。

$$s^2 = \frac{1}{n}\sum_{j=1}^{k}(x_j - \bar{x})^2 f_j, \qquad s^2 = \sum_{j=1}^{k}(x_j - \bar{x})^2 p_j \tag{4}$$

上のデータの分散を度数分布表から求める。

$$s^2 = \frac{(50-62.2)^2\cdot 1 + \cdots + (75-62.2)^2\cdot 2}{25} = 42.16$$

分散（2）　度数分布表から分散を求める場合も，公式 $s^2 = \overline{x^2} - \bar{x}^2$ が成立する。

$$s^2 = \frac{1}{n}\sum_{j=1}^{k} {x_j}^2 f_j - \bar{x}^2, \qquad s^2 = \sum_{j=1}^{k} {x_j}^2 p_j - \bar{x}^2 \tag{5}$$

上のデータの分散を度数分布表から求める。

$$s^2 = \frac{50^2\cdot 1 + 55^2\cdot 6 + \cdots + 75^2\cdot 2}{25} - 62.2^2 = 42.16$$

分散（3）　2 番目の公式はさらに変形できる。

$$s^2 = \frac{n\sum {x_j}^2 f_j - \left(\sum x_j f_j\right)^2}{n^2} \tag{6}$$

上のデータの分散を度数分布表から求める。

$$\begin{aligned} s^2 &= \frac{25\,(50^2\cdot 1 + 55^2\cdot 6 + \cdots + 75^2\cdot 2) - (50\cdot 1 + 55\cdot 6 + \cdots + 75\cdot 2)^2}{25^2} \\ &= \frac{25\cdot 97775 - 1555^2}{25^2} = 42.16 \end{aligned}$$

標準偏差　度数分布表から求める場合も，標準偏差は分散の平方根とする。上のデータの標準偏差は $s = \sqrt{42.16} = 6.49$ となる。

3.3　変数の変換

もとの変数 x から新しい変数 y を，$y_i = ax_i + b$ のように定めると，その平均値について，$\bar{y} = a\bar{x} + b$ が成り立つ。

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) = a\cdot\frac{1}{n}\sum_{i=1}^{n} x_i + b = a\bar{x} + b$$

また，分散について，$s_y^2 = a^2 s_x^2$ が成り立つ。

$$\begin{aligned} s_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}\{(ax_i + b) - (a\bar{x} + b)\}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\{a(x_i - \bar{x})\}^2 = a^2\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = a^2 s_x^2 \end{aligned}$$

$y_i = ax_i + b$ のとき，平均値，分散，標準偏差について，次のことが成り立つ。

$$\bar{y} = a\bar{x} + b, \qquad s_y^2 = a^2 s_x^2, \qquad s_y = |a| s_x \tag{7}$$

変数の変換により分散を求める　度数分布表から平均値や分散を求める場合，c に階級値のいずれか，d に階級の幅を代入すると y_j が簡単な値になる。

階級	階級値 x	y	度数	相対度数
$47.5 \sim 52.5$	50	-3	1	0.04
$52.5 \sim 57.5$	55	-2	6	0.24
$57.5 \sim 62.5$	60	-1	6	0.24
$62.5 \sim 67.5$	65	0	7	0.28
$67.5 \sim 72.5$	70	1	3	0.12
$72.5 \sim 77.5$	75	2	2	0.08

$c = 65$（最頻値），$d = 5$（階級の幅），$y_j = (x_j - 65)/5$ とおくと，平均値は，

$$\bar{y} = \frac{(-3)\cdot 1 + (-2)\cdot 6 + \cdots + 2\cdot 2}{25} = \frac{-14}{25} = -0.56$$
$$\bar{x} = 65 + 5\cdot\bar{y} = 60 + 5\cdot\frac{-14}{25} = 62.2$$

分散は，

$$s_y^2 = \frac{(-3)^2\cdot 1 + \cdots + 2^2\cdot 2}{25} - \bar{y}^2 = \frac{50}{25} - \left(\frac{-14}{25}\right)^2 = \frac{1054}{25^2} = 1.6864$$
$$s_x^2 = 5^2\cdot s_y^2 = 5^2\cdot\frac{1054}{25^2} = 42.16$$

3.4　標準化変数

標準化（standardizing）　観測値 $x_1, x_2, \cdots, x_n$ の平均を $\bar{x}$，標準偏差を s_x（ただし $s_x > 0$）とする。次のように定めた変数 z_i を標準化変数または基準化変数という。あるいは簡単に標準化，基準化という。

$$z_i = \frac{x_i - \bar{x}}{s_x} \tag{8}$$

このとき，$x_i = \bar{x} + s_x z_i$ と変形できるから，平均値は，$\bar{x} = \bar{x} + s_x\bar{z}$ より，$\bar{z} = 0$ となり，分散は，$s_x = |s_x|s_z$ より，$s_z = 1$ となる。

$$\bar{z} = 0, \qquad s_z = 1 \tag{9}$$

標準化すれば，平均値と標準偏差を一定値（0 と 1）にそろえることができるため，分布が異なる集団であっても比較できるようになる。

偏差値（T-score）　標準化変数 z_i から次のように定めた変数 t_i を偏差値という。[*ii] [*ii]

$$t_i = 50 + 10z_i = 50 + 10\cdot\frac{x_i - \bar{x}}{s_x}$$

変数の変換の公式から，偏差値の平均値は $\bar{t} = 50$，標準偏差は $s_t = 10$ となる。

[*ii] 偏差値と同種のものに知能指数等がある。$\mathrm{IQ} = 100 + 15z$

例　下表において 1 段目の変数 x の平均は $\bar{x} = 61$，標準偏差は $s_x = 13$ だから，標準化 z，偏差値 t は次のようになる。

x	38	46	53	57	59	64	65	66	80	82
z	-1.77	-1.15	-0.62	-0.31	-0.15	0.23	0.31	0.38	1.46	1.62
t	32.3	38.5	43.8	46.9	48.5	52.3	53.1	53.8	64.6	66.2

たとえば，$x_1 = 38$ に対して，標準化は，$z_1 = (38 - 61)/13 = -1.77$，偏差値は，$t_1 = 50 + 10 \times (-1.77) = 32.3$ となる。

3.5　平均偏差

中央値からの平均偏差（MD; mean absolute deviation）　偏差の絶対値を絶対偏差という。中央値からの絶対偏差の平均値のことを平均絶対偏差あるいは簡単に平均偏差という。$\tilde{x}$ は中央値を表す。

$$\mathrm{MD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \tilde{x}|$$

変動係数 CV は標準偏差を平均値で割ったものであった。平均偏差を中央値で割った値 $\mathrm{MD}/\tilde{x}$ は，変動係数と同じような目的で使うことができる。

平均値からの平均偏差　中央値からではなく，平均値からの絶対偏差の平均値を平均偏差とすることもある。$\bar{x}$ は平均値を表す。

$$\mathrm{MD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

3.6　歪度・尖度

分布の中心の位置は代表値によって，分布の広がりは散布度によって表される。分布の形状を表す量として，歪度や尖度が知られている。

歪度（skewness）　偏差の 3 乗の平均値を標準偏差の 3 乗で割った値のことを歪度（わいど）という。

$$\alpha_3 = \beta_3 = \frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{s}\right)^3$$

歪度が正のとき，右に歪んだ分布（峰が左にあり，裾が右に厚い分布）となり，歪度が負のときはその逆となる。対称な分布の歪度は $\alpha_3 = 0$ である。

尖度（kurtosis）　偏差の 4 乗の平均値を標準偏差の 4 乗で割った値 α_4 のことを尖度（せんど）という。

$$\alpha_4 = \frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{s}\right)^4$$

あるいは $\beta_4 = \alpha_4 - 3$ を尖度ということもある。

$$\beta_4 = \frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{s}\right)^4 - 3$$

尖度が大きいとき，峰が鋭く裾が厚い分布となり，尖度が小さいとき，峰が鈍く裾が薄い分布となる。正規分布の尖度は $\alpha_4 = 3$，$\beta_4 = 0$ である。

参考文献

統計学入門（基礎統計学）
東京大学教養学部統計学教室（編）東京大学出版会 978-4-13-042065-5

統計学
久保川 達也 （著） 東京大学出版会 978-4-13-062921-8

はじめての統計学
道家 暎幸（共著） コロナ社 978-4-339-06113-0

確率統計 新版 （新版数学シリーズ）
岡本 和夫 （ほか著） 実教出版 978-4-407-32171-5

統計学序論 改訂版
山本 義郎 （著） 東海大学出版部 978-4-486-02133-9

確率統計（高専テキストシリーズ）
上野 健爾（監修） 森北出版 978-4-627-05561-2

基本統計学 第４版
宮川 公男 （著） 有斐閣 978-4-641-16455-0

新統計入門
小寺 平治 （著） 裳華房 978-4-7853-1099-8

Schaum's Outline of Introduction to Probability and Statistics
Seymour Lipschutz （著） McGraw-Hill Education 978-0-07-176249-6

A Dictionary of Statistics
Graham Upton （著） Oxford Univ Pr 978-0-19-967918-8

www5e.biglobe.ne.jp/~emm386/statistics/