第 II 部　統計

3　散らばりの指標

ポイント

- 散らばりの指標の意味
- レンジ
- 四分位偏差
- 分散
- 標準偏差

3.1　はじめに

- 特長を言葉で表現する場合、独自な表現を理解してもらうのは困難です。
- その界隈の標準的な視点で標準的な表現で特徴をつたえるのが合理的です。

3.1.1 総和の定義と公式

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n \tag{1}$$

$$\sum_{i=1}^{n} c = nc \qquad ; c \text{ は定数} \tag{2}$$

$$\sum_{i=1}^{n} ax_i = a\sum_{i=1}^{n} x_i \qquad ; a \text{ は定数} \tag{3}$$

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i \tag{4}$$

3.1.2 平均の定義

x_i の総和を n で割った値を『平均』といい $\bar{x}$ であらわす。

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5}$$

従って

$$n\bar{x} = \sum_{i=1}^{n} x_i \tag{6}$$

である。

3.2 散らばりの指標

代表値は分布の位置を示す指標とみることができる。位置以外に分布を区別する指標として形状の指標を考える。分布の形状を表す指標は沢山あるがここでは『散らばり』の尺度を紹介する。位置の指標と散らばりの尺度の二つを用いれば大まかな分布の形状を記述できる。

- レンジ
- 四分位範囲
- 四分位偏差
- 平均偏差
- 分散
- 標準偏差

3.2.1 レンジ

最大値と最小値の差を『レンジ[1)]』という。観測値 x_i の最大値を $\max(x)$、最小値を $\min(x)$ とするとレンジは以下の式で定義される。

$$\text{レンジ} = \max(x) - \min(x) \tag{7}$$

レンジはもっとも単純な散らばりの指標であり、そもそも n 個あるデータのうち二つしか使っていないため、$n-2$ 個のデータの特徴を表現できない。また、とびぬけて大きな値や小さな値の影響を受けやすいといった特徴がある。

1) 『範囲』とも呼ばれるが、一般名詞と区別できないので本講座ではレンジと呼称する。

3.2.2 四分位範囲

第 3 四分位数と第 1 四分位数の差を『四分位範囲 (QR)』という。

$$QR = Q_3 - Q_1 \tag{8}$$

四分位数はデータを 4 等分する点である。第 3 四分位数は上位 25 %の点である。つまり下位 75 % の点である。第 1 四分位数は下位 25 %の点である。上位 25% をのぞき、そこから更に下位 25 % を除いた範囲が四分位範囲である。四分位範囲は、中央の 50 % のデータが収まる範囲である。

3.2.3 四分位偏差

四分位範囲の半分を『四分位偏差 (QD)』という。

$$QD = \frac{1}{2}(Q_3 - Q_1) \tag{9}$$

$$= \frac{1}{2}(Q_3 - Q_2 + Q_2 - Q_1) \tag{10}$$

$$= \frac{1}{2}((Q_3 - Q_2) + (Q_2 - Q_1)) \tag{11}$$

(11) 式カッコ内 第一項は第 3 四分位数 (Q_3) と中央値との隔たりであり同第二項は、第 1 四分位数 (Q_1) と中央値の隔たりである。第 1 四分位数と第 3 四分位数それぞれの中央値からの隔たりの平均が四分位偏差である。

3.2.4 平均からの偏差

観測値 (x_i) と観測値の平均 $(\bar{x})$ の差を『平均からの偏差』という。

$$x_i - \bar{x} \tag{12}$$

平均からの偏差は各観測値と平均の差であるため、$x_i > \bar{x}$ であれば正、$x_i < \bar{x}$ であれば負の符号を持っている。『平均からの偏差』の総和を求めると、0 になる。このように、それぞれの値が持つ符号によって値が打ち消しあうことを『符号の問題』という。

問題 II−3 − 1

次の値を計算しなさい

$$\sum_{i=1}^{n}(x_i - \bar{x})$$

解例 II−3 − 1

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x}) &= \sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\bar{x}\\ &= \sum_{i=1}^{n}x_i-n\bar{x}\\ &= \sum_{i=1}^{n}x_i-n\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\\ &= \sum_{i=1}^{n}x_i-\sum_{i=1}^{n}x_i\\ &= 0\end{aligned}$$

3.2.5 平均偏差

『平均からの偏差』の絶対値の平均を『平均偏差 (AD)』という。

$$AD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \tag{13}$$

『平均からの偏差』の総和は 0 になるため散らばりの尺度としては使えない。そこで符号の問題を解決するために絶対値をとっている。

3.2.6 分散

『平均からの偏差』の二乗の平均を『分散 (V_x)』という。

$$V_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{14}$$

『平均からの偏差』の総和は 0 になるため散らばりの尺度としては使えない。そこで符号の問題を解決するために二乗している。

問題 II−3 − 2

次の値を計算しなさい

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

解例 II−3 − 2

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \frac{1}{n}\sum_{i=1}^{n}\left({x_i}^2-2\bar{x}x_i+\bar{x}^2\right)=\frac{1}{n}\left(\sum_{i=1}^{n}{x_i}^2+\sum_{i=1}^{n}(-2\bar{x}x_i)+\sum_{i=1}^{n}\bar{x}^2\right)\\
&=\frac{1}{n}\left(\sum_{i=1}^{n}{x_i}^2-2\bar{x}\sum_{i=1}^{n}x_i+n\bar{x}^2\right)\\
&=\frac{1}{n}\sum_{i=1}^{n}{x_i}^2-2\bar{x}\,\frac{1}{n}\sum_{i=1}^{n}x_i+\frac{1}{n}n\bar{x}^2\\
&=\frac{1}{n}\sum_{i=1}^{n}{x_i}^2-2\bar{x}^2+\bar{x}^2\\
&=\frac{1}{n}\sum_{i=1}^{n}{x_i}^2-\bar{x}^2
\end{aligned}$$

3.2.7 標準偏差

『分散』の正の平方根を『標準偏差 (σ_x)』という。

$$\sigma_x = \sqrt{V_x} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{15}$$

問題 II−3 − 3

表 1 を用いて レンジ, 四分位偏差, 平均偏差, 分散, 標準偏差を求めなさい。

表 1　小学生の体重（サンプルデータ, 単位 kg）

20.5	25.5	22.5	25.5	29.0	21.5	24.0	25.5	22.5	25.5	25.5	27.5	29.0	31.5
22.5	25.5	25.5	31.5	24.0	27.5	30.5	29.0	34.0	26.5	30.5	24.0	24.0	26.5
29.0	29.0	26.5	29.0	31.5	22.5	25.5	25.5	29.0	31.5	25.5	29.0	30.5	31.5
29.0	30.5	29.0	32.5	30.5	29.0	27.5	30.5	32.5	31.5	32.5	34.0	29.0	27.5
27.5	25.5	30.5	29.0	31.5	30.5	30.5	32.5	34.0	31.5	30.5	32.5	31.5	32.5
36.5	37.5	26.5	30.5	25.5	27.5	30.5	31.5	32.5	32.5	34.0	29.0	30.5	31.5
30.5	36.5	31.5	32.5	30.5	32.5	37.5	29.0	31.5	32.5	34.0	31.5	27.5	30.5
32.5	31.5	32.5	30.5	31.5	32.5	32.5	30.5	35.5	36.5	32.5	36.5	36.5	37.5

問題 II−3 − 2

表 2 を用いて レンジ, 四分位偏差, 平均偏差, 分散, 標準偏差を求めなさい。

表 2 小学生の身長（サンプルデータ, 単位 cm）

112.5	120.0	127.5	115.0	117.5	125.0	117.5	120.0	130.0	125.0	125.0	132.5	135.0	122.5
120.0	125.0	127.5	125.0	127.5	127.5	117.5	125.0	120.0	127.5	120.0	130.0	127.5	130.0
125.0	130.0	130.0	130.0	127.5	132.5	122.5	127.5	130.0	127.5	132.5	135.0	132.5	135.0
137.5	117.5	122.5	122.5	125.0	125.0	120.0	125.0	122.5	125.0	132.5	125.0	122.5	125.0
127.5	132.5	130.0	132.5	127.5	130.0	137.5	130.0	132.5	137.5	135.0	140.0	145.0	120.0
125.0	132.5	130.0	135.0	137.5	122.5	130.0	132.5	140.0	135.0	130.0	135.0	132.5	137.5
140.0	137.5	137.5	132.5	125.0	137.5	125.0	130.0	132.5	137.5	135.0	140.0	127.5	132.5
140.0	132.5	140.0	135.0	137.5	132.5	135.0	137.5	142.5	135.0	140.0	142.5	147.5	150.0

解例 II−3 − 2

表 3 体重のばらつきの指標

レンジ	17.00
四分位偏差	2.50
平均偏差	2.93
分散	13.48
標準偏差	3.67

四分位数は Excel の QUARTILE.INC 関数を利用して算出。

解例 II−3 − 3

.

表 4 身長のばらつきの指標

レンジ	37.50
四分位偏差	5.00
平均偏差	5.72
分散	50.55
標準偏差	7.11

3.3 まとめ

- 代表値と散らばりの尺度によりデータの分布の大まかな把握ができる。
- レンジは最大値と最小値の差。
- 四分位範囲は第 3 四分位数と第 1 四分位数の差。
- 分散 (V_x) は『平均からの偏差』の二乗の平均。
- 標準偏差 (σ_x) は分散 (V_x) の正の平方根。
- 『平均からの偏差』と『平均偏差』は全く異なる意味を持つ用語。