

第 II 部 統計

2 代表値

ポイント

- 総和記号の定義と公式
- 代表値とは

- 算術平均・最頻値・中央値
- 四分位数
- 箱ひげ図

2.1 はじめに

- 視覚情報を言語化することは困難です。
- 言葉で説明するためには理由や意味を伝える必要があります。
- 定義は言語化されています。

2.2 総和

定義

n 個の数値があり、 x_i を i 番目の値とする。 n 個の x_i を全て足し合わせた値を『総和』とよび $\sum_{i=1}^n x_i$ という記号であらわす。

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

2.2.1 定数の総和

全ての x_i が定数 (c) であった場合、すなわち $\forall x_i = c$ のとき、以下の記号であらわす。

$$\sum_{i=1}^n c = c + c + \cdots + c \quad (2)$$

数値はそもそも n 個あるので

$$= nc \quad (3)$$

である。

2.2.2 定数倍の総和

全ての x_i を定数 (a) 倍した値の総和を以下の記号であらわす。

$$\sum_{i=1}^n ax_i = ax_1 + ax_2 + \cdots + ax_n \quad (4)$$

右辺をカッコに括り共通因子 (a) を掃きだすと

$$= a(x_1 + x_2 + \cdots + x_n) \quad (5)$$

カッコ内は (1) 式なので

$$= a \sum_{i=1}^n x_i \quad (6)$$

である。

2.2.3 和の総和

二つの数値の組が n 組あり、 (x_i, y_i) を i 番目の組とする。 x_i と y_i の和の総和を以下の記号であらわす。

$$\sum_{i=1}^n (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + \cdots + (x_n + y_n) \quad (7)$$

$$= x_1 + y_1 + x_2 + y_2 + \cdots + x_n + y_n \quad (8)$$

$$= (x_1 + x_2 + \cdots + x_n) + (y_1 + y_2 + \cdots + y_n) \quad (9)$$

$$= \sum_{i=1}^n x_i + \sum_{i=1}^n y_i \quad (10)$$

3 変数以上の場合も同様である。

2.2.4 総和の公式

総和の定義より以下の公式を導き出すことができる。

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

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

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

問題 II-2-1

次の式を計算しなさい。ただし、 a, c は定数とする。

$$\sum_{i=1}^n (ax_i + c)$$

解例 II-2-1

$$\begin{aligned}\sum_{i=1}^n (ax_i + c) &= \sum_{i=1}^n ax_i + \sum_{i=1}^n c \\ &= a \sum_{i=1}^n x_i + nc\end{aligned}$$

2.3 代表値とは

分布を代表する値の事を言う。代表値は数量的概念によるものを指し、計算したり操作したり伝達したりするための客観性がある事が特徴である。

以下の説明において観測値は n 個あり i 番目の観測値を x_i であらわす。

2.3.1 最大値・最小値

n 個ある観測値において最大の値を『最大値』といい $\max(x)$ であらわし、最小の値を『最小値』といい $\min(x)$ であらわす。最大値・最小値を持つ観測値はそれぞれ一つとは限らず、複数ある場合もある。

また、特別に最大値が大きすぎる、最小値が小さすぎると判断され異常値として除く場合がある。

2.3.2 算術平均

観測値の総和を観測値の総数で割った値を『算術平均』『相加平均』『平均値』または単に『平均』といい、 $\bar{x}$ であらわす。

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (11)$$

2.3.3 幾何平均

$\forall x_i > 0$ とする。観測値の総乗の n 乗根を『幾何平均』または『総乗平均』といい x_G であらわす。

$$x_G = \sqrt[n]{x_1 \times x_2 \times x_3 \times \cdots \times x_n} \quad (12)$$

平均倍率を求めるときは幾何平均を用いる。

2.3.4 調和平均

$\forall x_i > 0$ とする。観測値の逆数の算術平均の逆数を『調和平均』といい、 x_H であらわす。

$$\frac{1}{x_H} = \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \quad (13)$$

$$x_H = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} \quad (14)$$

2.3.5 調整平均

データの最大値 $\max(x)$ と最小値 $\min(x)$ を除き、残りの $n - 2$ 個のデータの算術平均を『調整平均』といい、 x_T であらわす。

$$x_T = \frac{1}{n-2} \left(\sum_{i=1}^n x_i - \max(x) - \min(x) \right) \quad (15)$$

上位と下位をそれぞれ $\alpha\%$ ずつ除外した場合は $\alpha\%$ 調整平均という。

2.3.6 中央値

観測値を昇順にならべ、中央に位置する値を『中央値・メディアン (median) 』という。並べ替えた観測値を小さいものから $x_1, x_2, x_3, \dots, x_n$ とすると、観測値の総数 (n) が奇数 ($n = 2m + 1$) の場合は、中央値は $m + 1$ 番目の観測値 (x_{m+1}) の値である。

$$\overbrace{x_1, x_2, \dots, x_m}^{m \text{ 個}}, x_{m+1}, \overbrace{x_{m+2}, x_{m+3}, \dots, x_n}^{m \text{ 個}} \quad (16)$$

観測値の総数 (n) が偶数 ($n = 2m$) の場合は、 m 番目 (x_m) と $m + 1$ 番目の観測値 (x_{m+1}) の平均 $\left(\frac{x_m + x_{m+1}}{2} \right)$ を中央値とする。

$$\overbrace{x_1, x_2, \dots, x_m}^{m \text{ 個}}, \overbrace{x_{m+1}, x_{m+2}, \dots, x_n}^{m \text{ 個}} \quad (17)$$

2.3.7 四分位数

観測値を昇順に並べ、四等分した時の分割点を四分位数といいそれぞれ、 Q_1, Q_2, Q_3 であらわす。

第1四分位数 (Q_1) は下位 25% を分ける分割点にある観測値であり第2四分位数 (Q_2) は中央値に一致する。そして第3四分位数 (Q_3) は上位 25% を分ける分割点にある観測値である。

第1四分位数・第3四分位数の算出方法には多様な解釈があり、指標として利用する場合は明確に定義する必要がある。

$n = 4m$ の場合

$$\underbrace{\overbrace{x_1, x_2, \dots, x_m}^{2m}, \underbrace{x_{m+1}, \dots, x_{2m}}_m, \underbrace{x_{2m+1}, \dots, x_{3m}}_m, \underbrace{x_{3m+1}, \dots, x_{4m}}_m}_{2m} \quad (18)$$

なので

$$Q_1 = \frac{x_m + x_{m+1}}{2} \quad (19)$$

$$Q_2 = \frac{x_{2m} + x_{2m+1}}{2} \quad (20)$$

$$Q_3 = \frac{x_{3m} + x_{3m+1}}{2} \quad (21)$$

$n = 4m + 1$ の場合

$$\overbrace{x_1, \dots, x_m, x_{m+1}, \dots, x_{2m}}^{2m}, x_{2m+1}, \overbrace{x_{2m+2}, \dots, x_{3m}, x_{3m+1}, \dots, x_{4m+1}}^{2m} \quad (22)$$

なので第2四分位数は中央値に等しく、

$$Q_2 = x_{2m+1} \quad (23)$$

である。

Q_1 の算出時において、 x_{2m+1} を含まなければ

$$\underbrace{\overbrace{x_1, \dots, x_m}^m, \overbrace{x_{m+1}, \dots, x_{2m}}^m}_{2m}, x_{2m+1}, \underbrace{\overbrace{x_{2m+2}, \dots, x_{3m+1}}^m, \overbrace{x_{3m+2}, \dots, x_{4m+1}}^m}_{2m} \quad (24)$$

なので、第1四分位数 (Q_1) は

$$Q_1 = \frac{x_m + x_{m+1}}{2} \quad (25)$$

である。一方で、 x_{2m+1} を含めて Q_1 を算出する場合、利用されるデータは

$$\overbrace{\underbrace{x_1, x_2, \dots, x_m}_m, x_{m+1}, \underbrace{x_{m+2}, \dots, x_{2m}, x_{2m+1}}_m}^{2m+1} \quad (26)$$

なので

$$Q_1 = x_{m+1} \quad (27)$$

である。

この場合、 x_{2m+1} は Q_3 を求めるときにも再度利用され

$$\overbrace{\underbrace{x_{2m+1}, x_{2m+2}, \dots, x_{3m}}_m, x_{3m+1}, \underbrace{x_{3m+2}, \dots, x_{4m+1}}_m}^{2m+1} \quad (28)$$

の中央値が Q_3 となる。

x_{2m+1} は Q_1, Q_3 の算出において 2 度利用される。

$n = 4m + 2$ の場合

$$\underbrace{\overbrace{x_1, \dots, x_m, x_{m+1}, x_{m+2}, \dots, x_{2m+1}}^{2m+1}}_m, \underbrace{\overbrace{x_{2m+2}, \dots, x_{3m+1}, \dots, x_{4m+2}}^{2m+1}}_m \quad (29)$$

なので

$$Q_1 = x_{m+1} \quad (30)$$

である。

$n = 4m + 3$ の場合

$$\underbrace{\overbrace{x_1, \dots, x_{m+1}}^m, \overbrace{x_{m+2}, \dots, x_{2m+2}}^m, x_{2m+3}, \dots, x_{3m+3}}_{2m+2}, \overbrace{x_{3m+4}, \dots, x_{4m+3}}^m \quad (31)$$

第1四分位数 Q_1 の算出において x_{m+2} を含まなければ

$$Q_1 = x_{m+1} \quad (32)$$

であり、 x_{m+2} を含めて算出した場合

$$Q_1 = \frac{x_{m+1} + x_{m+2}}{2} \quad (33)$$

である。

2.3.8 最頻値

度数分布表において最大の度数をもつ階級値を『最頻値・モード (mode) 』という。最大の度数を持つ階級が複数ある場合は、それぞれが最頻値である。

2.3.9 数値例

表 1 を用いて 平均、中央値、最頻値、最大値、最小値、四分位数を求めなさい。

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

表 2 体重の代表値

最大値	37.5 cm
最小値	20.5 cm
平均	29.8 cm
中央値（第 2 四分位数）	30.5 cm
第 1 四分位数	27.5 cm
第 3 四分位数	32.5 cm
最頻値	30.5 cm

問題 II-2-2

表 3 を用いて 平均、中央値、最頻値、最大値、最小値、四分位数を求めなさい。

表 3 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

表 4 身長代表値

最大値	150.0 cm
最小値	112.5 cm
平均	130.1 cm
中央値（第 2 四分位数）	130.0 cm
第 1 四分位数	125.0 cm
第 3 四分位数	135.0 cm
最頻値	125.0 cm
	132.5 cm

2.4 箱ひげ図

『箱ひげ図』とは最小値・第1四分位数・中央値・第3四分位数・最大値を使い、データのばらつきを表現する統計グラフである。第1四分位数から第3四分位数までの範囲に箱を描き最大値と最小値をひげの端であらわしたものである。

図1 箱ひげ図



2.5 まとめ

- 代表値を使うとデータの特長を記述可能。
 - 最大値・最小値
 - 平均
 - 中央値
 - 最頻値
 - 四分位数
- 代表値を並べるとデータの散らばり具合が把握できる。