

1.2 代表値

変数 x の n 個のデータの値が x_1, x_2, \dots, x_n とする。

- 平均値 $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$
- 中央値 (メジアン) データを大きさの順に並び替えたものを $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ とする。このとき、

$$\text{中央値} = \begin{cases} x_{(\frac{n+1}{2})} & n \text{ が奇数のとき} \\ \frac{1}{2} (x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & n \text{ が偶数のとき} \end{cases}$$

例題 1.1 次のデータの平均値と中央値を求めよ。

- (1) 42, 38, 40, 44, 52 (2) 42, 38, 40, 44, 52, 198

解答: (1) 平均値: $\bar{x} = \frac{42 + 38 + 40 + 44 + 52}{5} = 43.2$

中央値: データを大きさの順に並べると $38 < 40 < 42 < 44 < 52$ となるので、42。

(2) 平均値: $\bar{x} = \frac{42 + 38 + 40 + 44 + 52 + 198}{6} = 69$

中央値: $38 < 40 < 42 < 44 < 52 < 198$ となるので、 $\frac{1}{2}(42 + 44) = 43$. □

注意 平均値は数学的にいろいろよい性質をもっており、通常は平均値を用いることが多い。しかし、上記のように、平均値は他のデータからかけ離れた値をもつ「はずれ値」(次ページの箱ひげ図の5を参照のこと)の影響を受けやすいが、中央値はそうでないことがわかる。また、給与や貯蓄額のように指数的に変動すると考えられるデータの代表値として平均値を用いる場合には、注意が必要である。実際、平成22年国民生活基礎調査による所得金額階級別にみた世帯数のデータでは平均値549.6万円であり、中央値が438万円である。また、最頻値(度数が一番高い階級)は200-300万円である。^{*1}

1.3 散布度

- 分散 $s^2 = \frac{1}{n} \{ (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \}$
- 標準偏差 $s = \sqrt{\frac{1}{n} \{ (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \}}$

変数 x の測定単位が例えば「点」のとき、分散の単位は「点²」となってしまう。一方、標準偏差は変量と同じ測定単位となる。また、分散が0となるのはすべてのデータの値が一致するときに限ることに注意する。

n 個のデータの値 x_1, x_2, \dots, x_n を大きさの順に並び替えたものが $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ であった。

- 範囲 $x_{(n)} - x_{(1)}$ (データの最大値と最小値の差)

- 四分位数

$n = 2m$ が偶数のとき、

$x_{(1)}, x_{(2)}, \dots, x_{(m)}$ を下位のデータ、 $x_{(m+1)}, x_{(m+2)}, \dots, x_{(2m)}$ を上位のデータと、

$n = 2m + 1$ が奇数のとき、

$x_{(1)}, x_{(2)}, \dots, x_{(m)}$ を下位のデータ、 $x_{(m+2)}, x_{(m+3)}, \dots, x_{(2m+1)}$ を上位のデータという。

ここで、上位のデータ、下位のデータともに m 個のデータからなることに注意する。このとき、

第1分位数 Q_1 は下位のデータの中央値 第3分位数 Q_3 は上位のデータの中央値

と定める。なお、第2分位数 Q_2 はデータ全体の中央値(通常中央値)とする。

これを用いて、四分位範囲を $Q_3 - Q_1$ 、四分位偏差を $\frac{1}{2}(Q_3 - Q_1)$ と定める。

^{*1} <http://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa10/2-2.html>

この分布の様子は異様に思えるかもしれないが、所得の分布はこのような形状を取ることがよく知られている。

例題 1.2 次のデータの第 1 分位数 Q_1 と第 3 分位数 Q_3 を求めよ。

(1) 65, 70, 47, 78, 92, 65, 89, 95, 59, (2) 65, 70, 47, 78, 92, 65, 89, 95, 59, 73

解答: (1) データを小さいほうから並べると 47, 59, 65, 65, 70, 78, 89, 92, 95 であるから、下位のデータは 47, 59, 65, 65. よって、 $Q_1 = \frac{59+65}{2} = 62$. 同様に上位のデータは 78, 89, 92, 95 より $Q_3 = \frac{89+92}{2} = 90.5$.

(2) 順に並べると 47, 59, 65, 65, 70, 73, 78, 89, 92, 95 であるから、 $Q_1 = 65, Q_3 = 89$. □

例題 1.3 次の数値は、ある授業の 30 人の学生についてのテストの点数である。

| | | | | | | | |
|----|----|-----|----|----|----|-------------------------------|------------------------------------|
| 65 | 70 | 54 | 78 | 89 | 65 | これを度数分布表にまとめると次のようになった。 | |
| 28 | 93 | 100 | 68 | 88 | 26 | | 階級値 25 35 45 55 65 75 85 95 計 |
| 64 | 66 | 65 | 87 | 50 | 54 | | 度数 2 3 1 3 10 5 3 3 30 |
| 37 | 91 | 73 | 62 | 32 | 39 | ただし、21 点以上 30 点以下の階級値を 25 とし、 | |
| 56 | 80 | 65 | 78 | 75 | 70 | 他も同様に 35, 45, ..., とした。 | |

このとき、このデータの第 3 分位数 Q_3 を求めよ。ヒント: まずどの階級にあるかを考えよ。

解答: データ数が 30 だから上位のデータは 15 個であるので、 Q_3 は大きいほうから 8 番目のデータとなる。よって、階級値 75 の階級に属しており、その大きいほうから 2 番目のデータとなる。この階級に属するデータを抜き出すと 78, 73, 80, 78, 75 であるから、これを順に並べると 73, 75, 78, 78, 80 となるので、 $Q_3 = 78$. □

問 1.1 例題 1.3 のデータの第 1 四分位数 Q_1 と中央値 m を求めよ。(解答: $Q_1 = 54, m = 65.5$.)

注意 (1) 四分位数の定義は複数ある。上記で定義したものは一般に Q_1 は下側ヒンジ、 Q_3 は上側ヒンジと呼ばれている。例えば表計算ソフト Excel では、平面上の n 個の点 $(1, x_{(1)}), (2, x_{(2)}), \dots, (n, x_{(n)})$ を順に折れ

線で結んでできる関数 $y = f(t)$ 、即ち、 $f(t) = \begin{cases} x_{(t)}, & t \text{ が自然数} \\ ([t] - t)x_{([t])} + (t - [t])x_{([t]+1)}, & \text{それ以外} \end{cases}$ を用い、 $Q_q =$

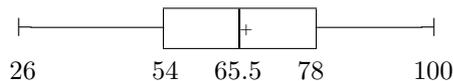
$f(1 + \frac{q}{4}(n - 1))$, $q = 1, 3$, と定めているようである。ここで、 $[t]$ は t 以上の最小の整数、 $\lfloor t \rfloor$ は t 以下の最大の整数を表す。この場合例題 1.2 の Q_3 は (1) $x_{(7)} = 89$, (2) $0.25x_{(7)} + 0.75x_{(8)} = 86.25$ となる。

(2) データの 最小値・第 1 四分位数・中央値・第 3 四分位数・最大値 を図にしたのが箱ひげ図である*2:

箱ひげ図は以下のように作成する。

- データの第 1 四分位点 Q_1 と第 3 四分位点 Q_3 により、全データの半数が含まれる箱を描く。
- 中央値 Q_2 を縦線で描く。
- 平均値を「+」で描く (省略されることもある)。
- 四分位範囲の 1.5 倍を箱の左右にとり、それを超えない内側のデータの最大値と最小値まで「ひげ」(左に「|—」, 右に「—|」) を引く。
- 内境界点の外側の左右に四分位範囲の 1.5 倍の長さを取り (外境界) その範囲にあるデータははずれ値として「○」でプロットする (全データの最小値と最大値まで「ひげ」を引く方法ではこれは描かない)。
- 外境界点の外側にあるデータを極値として「*」でプロットする (同上)。

例題 1.3 のデータの場合、平均値が 65.6, 最小値 26, 最大値 100 であるから、右ようになる。



ただし、平均値の数値は中央値に近いので記入しなかった。(この例でははずれ値は存在しない。)

*2 「稲葉芳成: 箱ひげ図について」を参考にした。新課程の数学 I の教科書では、4 で内側すべてに「ひげ」を引き、5 の「はずれ値」と 6 の「極値」を省力している。(この方法も一般的ですが、はずれ値を加える図も見かけます。)