第 1 章

データの整理と表現

もともと「統計」という言葉は，集めた多数のデータを整理して利用しようという実用的な目的のもとに使われるようになった．そのための手法を**記述統計学 (descriptive statistics)** と呼ぶ．この章では，多数のデータをどのように要約し，どのように表現するかを学ぶ．

1.1　データの集合から統計量を求める

100 人の男子高校生の体重を調べて，表 1.1 のような結果が得られた．

表 1.1　男子高校生 100 人の体重のデータ：単位は kg

43.6, 45.2, 45.4, 45.8, 47.2, 47.8, 48.2, 48.7, 48.8, 48.9, 49.0, 49.0, 49.4,
49.5, 49.8, 50.4, 50.5, 50.9, 50.9, 51.2, 51.2, 51.2, 51.3, 51.3, 51.6, 51.7,
51.7, 51.8, 52.0, 52.0, 52.1, 52.1, 52.1, 52.2, 52.3, 52.7, 52.7, 52.8, 52.9,
52.9, 53.1, 53.1, 53.8, 54.0, 54.5, 54.5, 54.6, 54.7, 54.7, 54.7, 54.8, 54.9,
55.1, 55.1, 55.2, 55.3, 55.4, 55.4, 55.4, 55.6, 55.7, 55.8, 55.9, 56.1, 56.3,
56.3, 56.3, 56.4, 56.5, 56.7, 56.8, 57.0, 57.1, 57.1, 57.2, 57.3, 57.6, 57.7,
57.8, 58.1, 58.4, 58.6, 58.7, 58.7, 58.7, 58.7, 59.1, 59.3, 59.9, 60.0, 60.1,
60.3, 60.5, 60.6, 60.6, 60.7, 61.3, 62.7, 64.2, 64.6

このようなデータの数値の並びをデータ列 $\boldsymbol{x}$ と呼び，次のように表現することにしよう．n はデータの数である．

$$\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\} \tag{1.1}$$

1.1.1　平均 $\bar{x}$, μ

これから**平均 (mean**[*1]**)** を求めるには，だれでも知っているように次のように計算すればよい．

$$\frac{1}{100}(43.6 + 45.2 + 45.4 + 45.8 + \cdots + 64.6) = 54.46$$

$\boldsymbol{x}$ の平均は $\bar{x}$ のように表記され，μ が使われることもある[*2]．平均は一般的に次のように定義される．

$$\begin{aligned} \bar{x} &= \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned} \tag{1.2}$$

総和記号 $\sum$ を使った書き方は短くて便利だが，ちょっとむつかしそうに見えるので，それを展開した形を思い浮かべて使うとよい．本書ではなるべく展開した形も併記する．

1.1.2　偏差

統計量そのものではないが，**偏差 (deviation)** もよく使われる量である．**平均偏差**とも呼ぶことがある．偏差は式 (1.3) で表されるように，あるデータが平均値からどれだけずれているかを意味する[*3]．

$$\delta x_i = x_i - \bar{x} \tag{1.3}$$

すべてのデータについての偏差の和はゼロになることが，次のようにして簡単に示せる．

$$\begin{aligned} \delta x_1 + \delta x_2 + \cdots + \delta x_n &= (x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) \\ &= (x_1 + x_2 + \cdots + x_n) - n\bar{x} \\ &= n \times \frac{1}{n}(x_1 + x_2 + \cdots + x_n) - n\bar{x} = 0 \end{aligned}$$

もっとも，平均よりも大きい分と小さい分が打ち消しあうので総和がゼロになると考えれば，式は見なくても直感的に理解できるだろう．

[*1] average もここで定義される平均の意味で使われるが，メジアン（後述）などデータの「真ん中」を表す他の尺度も含むあいまいな用語である．

[*2] μ はミューと読む．mean の m に相当するギリシャ文字である．

[*3] δ はデルタ．小さな差を表すのによく使われる．

1.1.3 分散 σ^2, 標準偏差 σ

データがどこを「中心」として分布しているのかを表すためには平均や後述するメジアンが使われる．それではデータがどの程度ばらばらに散っているかの目安としては，どのような量を考えればよいのだろうか．

偏差は，それぞれのデータの平均からのずれなので，すべての偏差を平均してみてその大きさで「ばらばら度」の尺度にしてみるという発想ではどうだろうか？しかし，上ですでに指摘したように，偏差の総和は常にゼロになるので，偏差の平均もゼロになってしまう．

そこで，偏差を 2 乗した値の平均として表される σ^2 という量を，データの広がりを表す尺度として定義する[*4]．

$$\begin{aligned} \sigma^2 &= \frac{1}{n}\left((x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \end{aligned} \tag{1.4}$$

σ^2 は**分散 (variance)** と呼ばれ，この値が大きいほどデータはばらばらに散っていることになる．

また，分散の平方根 σ は**標準偏差 (standard deviation)** [*5]と呼ぶ．

$$\sigma = \sqrt{\sigma^2} \tag{1.5}$$

■分散と標準偏差の使い分け

分散をデータの広がりの尺度として導入したが，どうしてわざわざその後で平方根をとった標準偏差というものを持ち込むのだろうか．

今，長さのデータを扱っているものとして，その単位が m であったとする．分散は 2 乗の平均だから，単位は m^2 ということになる．つまり分散はデータそのものとは異なった単位をもっているので，データや平均の値と比較することはできない．「10 m と 100 m^2 とどっちが大きい？」と聞かれても，答えるのは不可能だ．

そこで，分散の正の平方根である標準偏差を考えると，こちらはもとのデータと同じ単位をもっているので，たとえば平均の周りでデータがどのようにばらついているかを考えるには，標準偏差が有効だということになる．つまり，データから直接に計算できるのは分散なのだけれど，標準偏差のほうがデータと比較する尺度としては直観的に分かりやすいものだということになる．

[*4] σ はシグマと呼ぶ．

[*5] SD などと略されることがある．また，RMS (Root Mean Square) と呼ばれることもある．

以上を次のようにまとめておこう．

平均と標準偏差は，分布の中心と広がりをつかむためのワンセット

なお，よく似た概念として**標準誤差 (standard error)** があるが，それについては 84 ページで触れる．

■平均と分散はもっとも重要な統計量

データの集合の特徴を表す量のことを**代表値** (representative value / descriptive statistics) という．データの「真ん中」を代表する値には平均かメジアンが使われることが多いが，数学的には平均のほうがずっと扱いやすい．

そこで，平均をデータを代表する統計量，分散をデータのばらつきを表す基本的な統計量として取り扱うことが統計の中心的な作業になる．ただし，データの集団の実態とかあるいは実感といった見方からすると，次節で述べるメジアンや四分位数のほうが，より分かりやすい代表値であるということもしばしばある．

1.1.4　分散，標準偏差を求める別の公式

式 (1.4), 式 (1.5) は，別の形に導くことができ，そのほうが便利なことがある．すなわち，

$$\begin{aligned}
\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2-2\overline{x}x_i+\overline{x}^2) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+n\overline{x}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2-2n\overline{x}^2+n\overline{x}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2=\overline{x^2}-\overline{x}^2 \qquad (1.6)
\end{aligned}$$

最後の式に現われる $\overline{x^2}$ は $\frac{1}{n}(x_1^2+x_2^2+\ldots+x_n^2)$，つまり各データの **2 乗の平均**を意味している．

なおここで式の変形のために次の関係を利用した．

$$\sum_{i=1}^{n}\overline{x}x_i=\overline{x}x_1+\overline{x}x_2+\ldots+\overline{x}x_n=\overline{x}\sum_{i=1}^{n}x_i=\overline{x}\times n\overline{x}=n\overline{x}^2$$

$$\sum_{i=1}^{n} \overline{x}^2 = \overline{x}^2(\overbrace{1+1+\ldots+1}^{n}) = n\overline{x}^2$$

これらの式において，平均 $\overline{x}$ はデータ全体によってきまる定数だから，総和の記号の外にくくり出すことができることに注意しよう．

式 (1.6) は，次のきわめて大事な事実を教えてくれる．

分散 ＝二乗の平均 − 平均の二乗

この関係はしばしば利用される．また，分散を求めるための効率のよいアルゴリズムにもなっている．

問題 1–1 表 1.1 のデータから分散と標準偏差を求めよ．いずれも有効数字 4 桁で答えること．

問題 1–2 0 と 1 が合計 n 個あり，そのうちの 1 の割合が p であるようなデータを考える．このデータの平均と分散と標準偏差を求めよ．なおこの結果は，世論調査の結果の分析などで重要な意味を持つ．

♡ **標準偏差とデータのまとまり — チェビシェフの不等式** ♡

標準偏差がデータのばらつきの尺度であることはすでに説明したが，これについては**チェビシェフの不等式 (Chebyshev's inequality)** と呼ばれる有名な公式がある．数式を使わずに表現するとこうなる．

あるデータの集合の平均 μ と 標準偏差 σ が分かっているとする．その時，全体のうち $\mu \pm a\sigma$ の範囲からはみ出すデータの割合は，任意の a について $\frac{1}{a^2}$ 以下しかない．

たとえば，表 1.1 のデータでは，$\mu = 54.46$, $\sigma = 4.22$ となることが計算してみて分かる (問題 1–1)．そこで $a = 2$ ととってみると，平均の $\pm 2 \times 4.22$ の範囲は $54.46 - 2 \times 4.22 = 46.02$ と $54.46 + 2 \times 4.22 = 62.90$ を両端とする区間だ．定理が教えるのは，この範囲の外には，全部で 100 個あるうちたかだか $1/2^2 = 1/4$ 以下しかないということだ．つまり 25 個以下ということだ．一方，表を見てこの範囲から外れるデータの数を数えると全部で 6 個だから，チェビシェフの不等式と合致している．

こうやって実際に計算してみると，この不等式による「縛り」は緩すぎて，大してありがたくないように思えるかもしれない．しかしこの定理は，データは平均から遠ざかるほど割合が減少し，その減り方は標準偏差で測られるということを教えてくれるという意味で大切なものである．

★

1.1.5　メジアンと四分位数, median / quartile

■四分位数

データを同数に 4 等分したときに，全体の 1/4, 2/4, 3/4 の位置に相当する値を**四分位数 (quartile)** といい，3 つの値の小さい方から**第 1 四分位数 (first quartile)**，**第 2 四分位数 (second quartile)**，**第 3 四分位数 (third quartile)** という[*6]．ただし第 2 四分位数は次に出てくるメジアンに等しいので，四分位数は第 1 と第 3 についてのみ使うことが多い．これらの正確な計算法は次で述べる．

なお，一般にデータを任意に n 等分した三分位数，五分位数なども考えることができるが，最もよく用いられるのは四分位数である．

■メジアン

すべてのデータを大きさの順に並べた時に，中央に位置するデータの値を**メジアン (median)** または**中央値**という．メジアンは第 2 四分位数であり，平均と同様にデータの集合を代表する最も重要な統計量のひとつである．

■四分位数を計算して求める

データを 4 分割するといっても，データの数 n によって分割の仕方が変わるので，その場合によって計算の仕方が異なることになる．そこで，n を $4m$, $4m+1$, $4m+2$, $4m+3$ $(m = 0, 1, 2, \ldots)$ のように場合分けして考える．

図 1.1 を見てほしい．図中の $x_1, x_2, \ldots, x_n$ は昇順に並べられたデータの値だ．**これらは実際にはばらばらな値をとっているのだが，このように等間隔に配置して計算を進める．**データの数 n を 12 から 15 までと，およびそれらを一般化した $4m$ から $4m+3$ までの 4 通りの場合に分けて，上から順にデータ列の並びを示してある．

この図を使って実際に計算をする段取りは，次のようになる．

1. データの数 n の値によって，使うべき場合を決める．ここでは仮に 12 としよう．すると一番上の $4m$ の場合で行くことになる．
2. $\boldsymbol{Q_1}$ を決める点は，左から $n/4$ 番目と $n/4+1$ 番目，つまり x_3 と x_4 である．
3. 図から $\boldsymbol{Q_1}$ は x_3 と x_4 を 3 : 1 に内分する点だ．したがって次の式で求められる．

$$\frac{1}{4}(x_3 + 3 \times x_4)$$

4. 次に，$\boldsymbol{M}$ を決める点は $n/2 = 6$ 番目と $n/2+1 = 7$ 番目になる．ただし今度は 2

[*6] 1/4 分位数，2/4 分位数，3/4 分位数という呼び方もある．

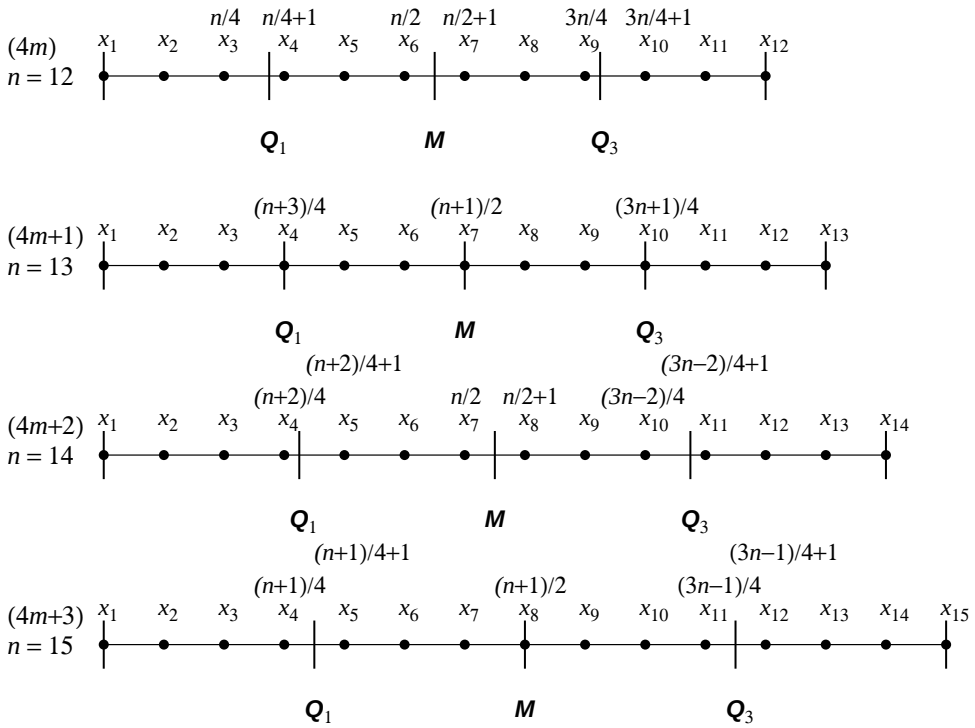


図 1.1　メジアン ($\boldsymbol{M}$), 第 1 四分位数 ($\boldsymbol{Q_1}$), 第 3 四分位数 ($\boldsymbol{Q_3}$) を計算するための場合分けと各分位数の位置．細かい意味は本文を参照のこと．

つの点を等分に内分しているので，次の式で求められる．

$$\frac{1}{2}(x_6 + x_7)$$

5. 最後に，$\boldsymbol{Q_3}$ を決める点は，$3n/4 = 9$ 番目と $3n/4 + 1 = 10$ 番目になる．今度はこれらを 1 : 3 に内分しているので，次の式で求められる．

$$\frac{1}{4}(3 \times x_9 + x_{10})$$

例題 1–1　メジアンと四分位数を求める

表 1.1 のデータから，第 1 四分位数 $\boldsymbol{Q_1}$，メジアン $\boldsymbol{M}$ ，第 3 四分位数 $\boldsymbol{Q_3}$ を求めよ．

図 1.1 を使ったデータの数 $n = 100$ は $4m$ の場合になるから．図を参考にして計算に使う各点の値を決めると次のようになる．

$$n/4 = 25,\ n/4 + 1 = 26,\ n/2 = 50,\ n/2 + 1 = 51,\ 3n/4 = 75,\ 3n/4 + 1 = 76$$

これから次のように計算して結果が得られる．

$\boldsymbol{M} = \frac{1}{2}(x_{50} + x_{51}) = \frac{1}{2}(54.7 + 54.8) = 54.75$
$\boldsymbol{Q_1} = \frac{1}{4}(x_{25} + 3x_{26}) = \frac{1}{4}(51.6 + 3 \times 51.7) = 51.675$
$\boldsymbol{Q_3} = \frac{1}{4}(3x_{75} + x_{76}) = \frac{1}{4}(3 \times 57.2 + 57.3) = 57.225$

例題 1–2　次のデータ列について，第 1 四分位数 $\boldsymbol{Q_1}$，メジアン $\boldsymbol{M}$ ，第 3 四分位数 $\boldsymbol{Q_3}$ を求めよ．解答は (　) 内に記した．

1. $\{3.2, 4.8, 14.0, 17.2, 22.8\}$
   (4.8, 14.0, 17.2)
2. $\{20.5, 30.5, 39.0, 46.5, 57.5, 59.0, 70.5, 80.5\}$
   (36.875, 52.0, 61.875)
3. $\{10.1, 10.7, 10.8, 11.2, 11.8, 12.5, 12.5, 12.8, 13.3, 13.8, 14.0, 14.7, 15.5, 16.3\}$
   (11.35, 12.65, 13.95)
4. $\{80.0, 80.0, 88.0, 92.8, 100.0, 108.8, 118.4, 129.6, 136.0, 144.8,$
   $146.4, 161.6, 176.0, 185.6, 192.0\}$
   (96.4, 129.6, 154.0)

1.1.6　四分位数と関係する用語

■パーセンタイル

四分位数ではデータを 4 つに分割する境目を考えるが．データを 100 分割して，100 分位数に相当する概念も**パーセンタイル (percentile)** と呼ばれてしばしば使われる．四分位数との関係では，第 1 四分位数が 25 パーセンタイル，メジアンが 50 パーセンタイル，第 3 四分位数が 75 パーセンタイルに相当する．

■ヒンジ

四分位数 $\boldsymbol{Q_1}, \boldsymbol{Q_3}$ を求める手順はやや面倒なので，**ヒンジ** と呼ばれる値が使われることもある．この場合にも，次のように下側と上側の 2 つのヒンジがあり，それぞれ $\boldsymbol{Q_1}, \boldsymbol{Q_3}$ と近似的に一致する．

下側ヒンジ (lower hinge)　メジアン以下のデータのメジアンを指す．
上側ヒンジ (upper hinge)　メジアン以上のデータのメジアンを指す．

$\boldsymbol{x} = \{1, 2, 3, 4\}$ の場合, 下側ヒンジ $= 1.5$, 上側ヒンジ $= 3.5$ であり，$\boldsymbol{x} = \{1, 2, 3, 4, 5\}$ の場合, 下側ヒンジ $= 2$, 上側ヒンジ $= 4$ となる．データ数が偶数の場合，メジアンはデータ点に含まれないので，メジアンよりも小さいデータを使って下側ヒンジを求めてい

る．上側についても同様．

■五数要約，箱ひげ図

データの最小値，第 1 四分位数，メジアン，第 3 四分位数，最大値の 5 つをまとめて，**五数要約 (five number summary)** と呼ぶ．これによって，データ全体の幅，中央，全体の半数が入っている領域をつかむことができる．なお，五数要約の定義として**第 1 四分位数と第 3 四分位数の代わりに下側ヒンジと上側ヒンジを使うこともある**．いずれにしても大きな違いは出ないので，実際上の不都合はない．

表 1.1 のデータについては，すでに例題 1–1 で第 1 四分位数，メジアン，第 3 四分位数が求めてあるので，それに最小値と最大値を付け加えて，五数要約は次のようになる．

$$43.6,\ 51.675,\ 54.75,\ 57.225,\ 64.6$$

■箱ひげ図

五数要約をグラフィカルに表した**箱ひげ図 (box and whiskers plot, box plot)** がしばしば用いられる．図 1.1.6 に，代表的な箱ひげ図の形とその各部の意味を示した．数値は表 1.1 のデータを用いている．箱ひげ図を使うと，データ集合の分布の様子が視覚的によく分かる．

なお．箱ひげ図の形や表現する内容は統一されてはおらず，形や向きを変えたり，後述する外れ値を表示するなど，使う目的とセンスによってさまざまな描き方がある．

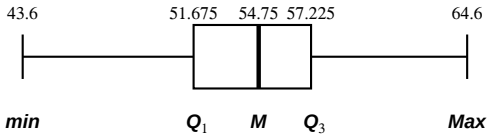


図 1.2　箱ひげ図．第 1 四分位数 $\boldsymbol{Q_1}$，メジアン $\boldsymbol{M}$，第 3 四分位数 $\boldsymbol{Q_3}$，を箱で表し，両端の「ひげ」で最小値 ***min*** 最大値 ***Max*** の位置を表す．

■四分位範囲 (IQR)

第 3 四分位数から第 1 四分位数を引いた値を**四分位範囲 (IQR**[*7]**)** といい，その半分の値を**四分位偏差**という．データの半数が含まれる幅を意味する量である．

[*7] Interquartile Range の略．

■外れ値

集団から遠く離れたデータのことを**外れ値 (outlier)** という．外れ値についての一致した数学的定義はなく，いくつかの基準が提唱されている．その中では，四分位数と関連付けた外れ値の定義[*8]がわかりやすく，次のように定義される．

データ x は次の条件のいずれかを満たすときに外れ値という．

$$x < \boldsymbol{Q_1} - k(\boldsymbol{Q_3} - \boldsymbol{Q_1}) \text{ または } x > \boldsymbol{Q_3} + k(\boldsymbol{Q_3} - \boldsymbol{Q_1})$$

言い換えれば，$\boldsymbol{Q_3} - \boldsymbol{Q_1} = \mathbf{IQR}$ だから，データが第 1 四分位数あるいは第 3 四分位数の外側に $\mathbf{IQR}$ の k 倍よりも遠く離れているときに外れ値と定義している． ここで k は必要に応じて 1.5〜3 ととる．

1.1.7　メジアンや分位数は頑健な代表値

■お年玉の金額の分布から

図 1.3 は，小学生 25 人がもらったお年玉の仮想的なデータを使って作った箱ひげ図である．一応現実的なデータに合わせるために現実の調査データを参考にしてある[*9]．計算に使ったデータは下の通りだ (単位 100 円)，

87, 143, 149, 163, 180, 186, 186, 212, 222,
247, 251, 255, 257, 261, 271, 274, 277, 281,
287, 296, 306, 347, 406, 449, 1300

平均では約 2 万 9 千円のところ，13 万円ももらった小学 2 年生がひとり含まれている．やはり子どもの世界でも，お金に関してはごく少数の「持てる者」が突出した金額を手にしているようだ．それをそのままプロットして見たのが，左の **A** である．見てのとおり，極端な外れ値が現れている．

この外れ値をいじって，2 番めの最大値と同じ程度にしてみたのが **B** だ．箱ひげ図を見ると，メジアンも第 1，第 3 四分位数も変化していない．

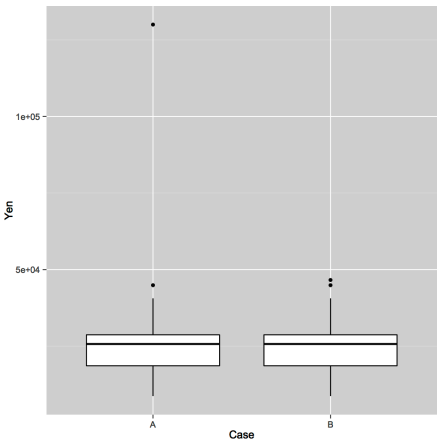


図 1.3　ある市の子どものお年玉の金額の分布を表した箱ひげ図 **A:**大きな外れ値あり，**B:**外れ値を修正してみたもの．

[*8] http://people.richland.edu/james/lecture/m170/を参照.
[*9] 川崎信用金庫「お年玉とお正月調査について」(2012)

こんどは，**A** と **B** の平均と標準偏差を比較してみよう．すると，平均値は 29,172 円から 25,836 円へと 3,300 円も下がり，標準偏差は 22,027 円から 8,920 円へと大幅に縮小している．このように外れ値 1 個のために，平均も標準偏差も少なからぬ影響を受けることが分かる．

このように，平均や分散は，大きな外れ値の存在によって敏感に変動する性質をもっている．一方，メジアンや四分位数は外れ値があっても，あまり，場合によっては全く，動かないことが分かる．このように「鈍感」であることを**頑健 (robust)** であると表現することがある．英語の読みのままで**ロバスト**ということもしばしばある．

1.1.8 残る命は何年だろうか

たとえば，ある病気にかかって手術を受けた人がいたとして，予後を知るために医学的な統計データを見たとしよう．データの中には，手術後の生存期間の情報をまとめたものもある．この人が頼りにするべきは，生存期間の平均だろうか，それともメジアンだろうか？

この治療の後で，かなりの人が 10 年程度生存し，15 年，20 年と生きた人もいたとしよう．しかし，2 割の人は 1 年以内に亡くなったものとする．すると余命の平均は約 8 年程度だが，メジアンの方は 12 年というといったケースが起こりうることになる．

こんな状況でこの人はどのように判断するのが賢明だろう？平均よりもメジアンを目安に考えるほうがよいのではないだろうか．「治療後のケアに十分な注意を払って，短命に終わることを避ければ．メジアンのところまでは行けそうだ」—そう考えることと，平均値を見て「あと 8 年の命か」と考えることとを比較すれば，このことは理解できるだろう．

こんなふうに，メジアンは「全体の真ん中あたり」という，いわば「並み」のポジションを表現しているものと考えられる．この後で扱う度数分布においては，このことがさらにはっきりと現れることになる．

1.2　度数分布

1.2.1　度数分布でデータを表す

生の数値を並べただけでは，これらのデータのもつ特徴をそこから直観的に見てとることは難しい．そこで，この種のデータを整理するために，**度数分布表 (frequency distribution table)** がしばしば使われる．度数分布表は，個々の数値を**表 1.2** のように**階級 (class)** に分けて，その**度数 (frequency)** を示したものである．度数は**頻度**ともいう．

また，ある階級までの度数の和の累計を**累積度数**という．

表 1.2　100 人の体重の統計を表す度数分布．表 1.1 のデータを使って構成した．

階級	階級値 (x_i)	度数 (f_i)	累積度数 (F_i)
43.0 – 45.0	44.0	1	1
45.0 – 47.0	46.0	3	4
47.0 – 49.0	48.0	6	10
49.0 – 51.0	50.0	9	19
51.0 – 53.0	52.0	21	40
53.0 – 55.0	54.0	12	52
55.0 – 57.0	56.0	19	71
57.0 – 59.0	58.0	15	86
59.0 – 61.0	60.0	10	96
61.0 – 63.0	62.0	2	98
63.0 – 65.0	64.0	2	100

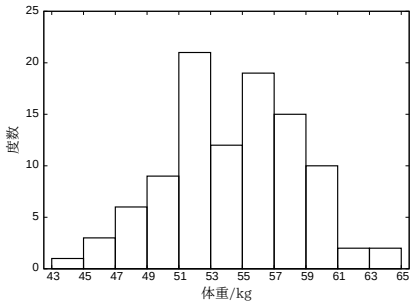


図 1.4　100 人の体重の統計を表すヒストグラム

また，度数分布をグラフで表して視覚的に把握しやすくしたものを**ヒストグラム (histogram)** という．表 1.2 の度数分布からは，図 1.4 の形のヒストグラムが作れる．

度数分布の表やヒストグラムを見ると，この集団の統計的な特徴を大づかみに見て取ることができる．すなわち，このデータによれば，中央付近の階級が大きな度数を持つ分布であり，平均はおよそ 53 から 55 の間に入るのではないかというふうに一目で推測できる．

1.2.2 度数分布から統計量を求める

度数分布表は集団のすべてのメンバーから得たデータを区分によって縮約したものである．その過程でいくらか情報量は失われるが，平均，メジアン，分散（と標準偏差）は，ほぼ正確に求めることが出来る．以下でその方法を考えよう．

■平均

度数分布から平均を求めるにはどうしたらよいだろうか．**表 1.2** を見てみよう．まず，体重の和は次のようにばらして書けることに注意する．

$$\begin{aligned} \text{総体重} &= \overbrace{44.0}^{1} + \overbrace{46.0 + 46.0 + 46.0}^{3} \\ &+ \overbrace{48.0 + 48.0 + 48.0 + 48.0 + 48.0 + 48.0}^{6} + \cdots \end{aligned} \tag{1.7}$$

これから同じ階級値の数値をまとめてやると，平均値は次のようにして計算できる．

$$\frac{\text{総体重}}{\text{総人数}} = \frac{44.0 \times 1 + 46.0 \times 3 + \ldots + 64.0 \times 2}{1 + 3 + \ldots + 2} = \frac{5446}{100} = 54.46 \tag{1.8}$$

式 (1.8) で得られる平均値は，個別のデータではなくて，階級という「塊」にまとめたものを使っているのであるから，幾分かの誤差を含むはずである．しかし多くのデータを扱う場合には，誤差は打ち消しあって十分に小さくなるので，ほぼ正しい平均値が得られる．

ここで式 (1.8) を一般化しておこう．データは k 個の階級に分けられており，階級値を $x_1, x_2, ..., x_k$，その度数を $f_1, f_2, ..., f_k$ とする[*10]．すると，上の例にならって，平均値を次のように表すことができる．

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i f_i = \sum_{i=1}^{k}(x_i \times \frac{f_i}{n}) \tag{1.9}$$

[*10] ここでは x_i が個々のデータの値ではなく，階級値であることに注意．

ここで n は $\sum_{i=1}^{k} f_i$，つまりデータの総数である．式 (1.9) は平均を表す式を 2 通りに表現したもので，2 つ目の表現は次の形をしていることに注意してほしい．確率分布でもこれによく似た形のものが表れる．

$$\text{平均} = (i \text{ 番目の階級値} \times i \text{ 番目の階級の割合}) \text{ の和}$$

■分散と標準偏差

分散は，式 (1.4) の定義を使えば次のようになる．

$$\sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 \times f_i}{n} = \sum_{i=1}^{k}(x_i - \overline{x})^2 \times \frac{f_i}{n} \tag{1.10}$$

2 番目の表現はやはり次の形をしている．

$$\text{分散} = (\ i \text{ 番目の階級の偏差の 2 乗} \times i \text{ 番目の階級の割合}) \text{ の和}$$

ここでも，分散の計算については，1.1.4 節で扱った場合と同様にして，2 乗の平均から平均の 2 乗を引けば求められる．

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i x_i^2}{n} - \overline{x}^2 \tag{1.11}$$

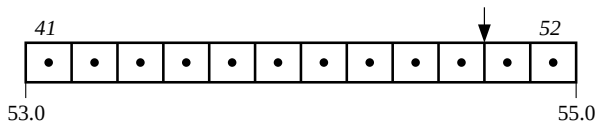
問題 1–3　式 (1.11) を利用して，表 1.2 のデータから体重の分散を求めなさい．

■メジアン

度数分布表から**メジアン**を求めるにはどうしたらよいだろうか．そのためには，ちょうど中央に位置する人の体重（総数が偶数の場合には中央の二人の体重の中間）を推定すればよい．この場合には 50 人目と 51 人目の人のデータの中間を推定したい．

累積度数を目安にして表を見ていくと，階級 (53.0–55.0) に 41 人目から 52 人目までの 12 人がいることがわかる．つまり，53.0 と 55.0 を両端とする区間の中に，12 人が並んでいるわけである．この並び方は等間隔ではないが，仮に等間隔と仮定して計算すればよい．

下のようにこれらの 12 人を並べたとすると，下の図のように考えて，50 人目と 51 人目の人の境目の位置は次の式で計算できる．



$$53.0 + \frac{2.0}{12} \times 10 = 54.666...$$

こうしてメジアンとして 54.67 が得られた．この値は真の値 54.75 にかなり近い．

■モード

この度数分布では，最も多くの度数をもつ階級の階級値は 52.0 である．このとき，この分布の**モード (mode)** あるいは**最頻値**は 52.0 であるという．モードも集団の代表値のひとつである．

1.3　平均，メジアン，モード，どれが全体を代表するのか

多数のデータの代表値としてもっともよく使われる統計量はいうまでもなく平均である．私たちの目に触れる社会の統計データでメジアンやモードという言葉が使われることはめったにない．それでは平均というのは，その集団を代表するパラメータとして他の 2 つの代表値よりも「優れて」いるのだろうか．

1.3.1　平均所得は「ふつうの所得」を意味するか

図 1.5 は，総理府が毎年実施している国民生活基礎調査のレポートから，各世帯ごとの所得金額の分布のヒストグラムを引用したものである．

これを見ると全世帯の平均所得は 538 万円であり，メジアン (中央値) はそれより 100 万円以上低い 427 万円となっている．モードは書かれていないが，所得が 300–400 万円の階級が最も多いので，階級値をとって約 350 万円とみなすことができる．つまり，この所得分布の代表値をみると，モードがもっとも小さく，その上にメジアンが来て，そして平均が最も大きな値をとっている．

この状況をまとめると次のようになるだろう．

- モードから判断すると，年間所得が 300 万円台の階層が最も多数を占める．
- メジアンから判断すると，真ん中の世帯は 427 万円の所得を得ている．
- 低所得者層から高所得層までを平均した所得は 538 万円である．

モードはもっともありふれた階層とみなせるから，これは「ふつうの世帯」としてランキングされるとも言える．またメジアンは，いうまでもなく「真ん中」だ．アンケートなどによく出てくる「中の中」という人々にあたるといってもよい．ところが平均はこれらよりもかなり高い所得のところにあるわけで，国民の実感とはかなりずれているわけだ．このようなことは，平均よりもずっと離れたデータの方が，平均に近いデータよりも，平均値に強く影響を与えることから起きる．それについて，少し考えておこう．

仮に図 1.5 の分布に，年間所得が平均よりも 100 万円多い 638 万円の世帯が 0.1% 加わった場合を考えてみよう．それによって平均がどうなるかは，次のように簡単に計算で

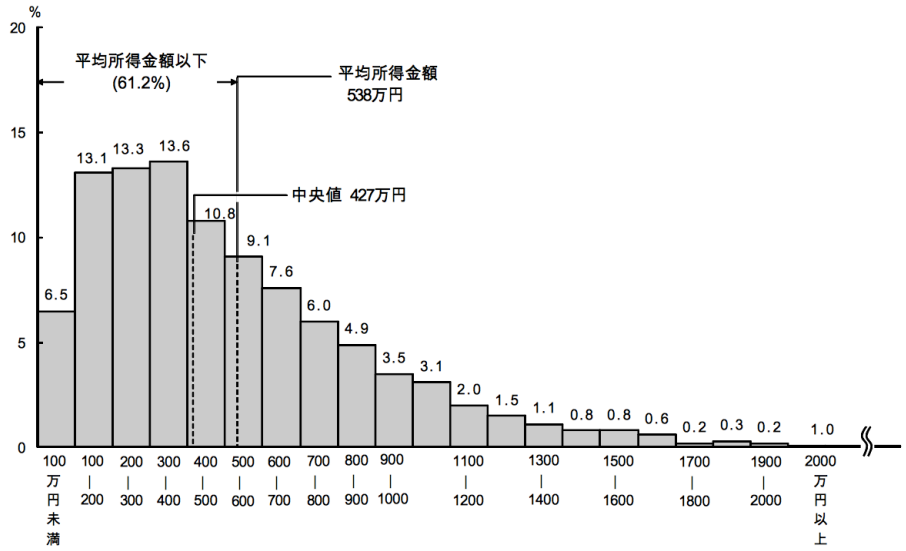


図 1.5　世帯ごとの所得金額の分布：平成 23 年度国民生活基礎調査

きる．

$$538 \times 0.999 + 638 \times 0.001 = 538.1$$

単位は 1 万円としてある．結果，1000 円だけ上昇するわけだ．今度は，年間所得が平均よりも 1 億円多い 10,538 万円の世帯が 0.1% 加わった場合で同じ計算をしてみると，548.1 万円，つまり 10 万円上昇することになる．

国民の所得は，最低のゼロから始まって上は何億円という人もいるはずだ[*11]．ということは，平均よりも桁外れに多い所得を得ている人が少数だけ存在しているために，実感とはかけ離れた平均所得が統計に現れることになる．

このように，この種の統計では得てして平均の値がさも一般庶民のものであるかのようにメディアでも政治でも扱われがちだが，それは実は虚像であって，モードやメジアンの方が，国民の実態を反映した代表値であると考えるのが妥当であろう．

[*11] 筆者はそのへんの想像力もお金の知識もないので，貧しい想像で書いている．

1.3.2　試験の成績の分布は

表 1.3 は国立教育政策研究所の平成 14 年度高等学校教育課程実施状況調査[*12] から引用した学力試験の結果の一部である．

まず国語の成績について平均点，メジアン，モードを度数分布から知ることができ，それぞれ 15.9, 17, 19 となっている．この試験では，23 点満点に近い成績が多かったために，平均は下位の得点に引っ張られることになってしまい，真ん中に位置する点数（メジアン），仲間がいちばん多い点数（モード）よりもかなり平均点が低くなっている．ここ例でも所得の統計の時と同様に，平均の値がどちらか一方に大きくずれた階級があるために，引きずられてしまうということが起きている．

一方，数学の成績について平均点，メジアン，モードを求めると，それぞれ 8.0, 8, 15 となっていて，モードが極端な値をとってしまっている．これは数学のやさしい試験で起こりがちな結果で，一応わかっている人は満点や高得点を取れるが，かなりの数の数学を不得手とする階層がいるために，低得点の側にもうひとつ山ができるのである．このような時にはモードはほとんど意味をなさない．メジアンは安定して真ん中の人がどの辺なのかを示している．このことは図 1.6 のようにデータの分布をグラフにしてみるとよくわかる．

表 1.3　全国の高校生対象の学力試験 (2012 年) における国語と数学の得点分布

	国語		数学	
素点	人数	累計	人数	累計
0	54	54	420	420
1	14	68	730	1150
2	27	95	983	2133
3	52	147	1118	3251
4	76	223	1122	4373
5	118	341	1045	5418
6	149	490	950	6368
7	169	6 59	993	7361
8	286	945	948	8309
9	343	1288	889	9198
10	424	1712	937	10135
11	518	2230	907	11042
12	678	2908	873	11915
13	737	3645	943	12858
14	927	4572	999	13857
15	1023	5595	1551	15408
16	1171	6766	–	–
17	1268	8034	–	–
18	1426	9460	–	–
19	1450	10910	–	–
20	1278	12188	–	–
21	1030	13218	–	–
22	569	13787	–	–
23	151	13938	–	–

以上のように，平均は，そこから遠く外れたデータからの影響を受けやすいが，メジアンは「真ん中」がどこかということを的確に示すという意味では，「よい代表値」であると言える．とはいえ，このような分布を得た場合に教育関係者や政治家が考えるべきことは，数学がきわめて苦手な生徒たちの大きな集団が我が国の学校に存在することを認識して，どのような方策をとるべきかということだろう．ヒストグラムの形は，統計的代表値では表しきれない情報を含むことがしばしばある．

問題 1–4　ある幼稚園の構成人員の年齢を調べたところ次のようになっていた．平均とメ

[*12] http://www.nier.go.jp/kaihatsu/katei_h14/index.htm

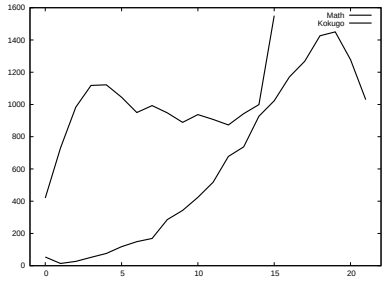


図 1.6　表 1.3 のデータから作成した得点分布.

ジアンを求めて，これら 2 つの代表値の違いについて考えなさい.

年齢	3	4	5	6	22	25	46	49	70	75
人数	15	28	31	15	1	1	1	1	1	1