データの平均値は，式 (4.1) のような計算で求めることもできるが，階級値とその度数を用いて次のように求めることもできる．

n 個の変量からなるデータの階級値が $y_1, y_2, \cdots, y_k$ であり，階級値 y_i $(i = 1, 2, \cdots, k)$ に対する度数が f_i であるとする[3]．このときこのデータの平均値 $\overline{x}$ は，

$$\overline{x} = \frac{1}{n}(y_1 f_1 + y_2 f_2 + \cdots + y_k f_k)$$

である．

注

例えば，あるクラスでテストを行った結果，60 点の人が 3 人，70 点の人が 4 人，80 点の人が 3 人いた場合，テストの点の平均値 $\overline{x}$ は，式 (4.1) にならうと，

$$\overline{x} = \frac{60+60+60+70+70+70+70+80+80+80}{10} = 70$$

のように求めることもできる．一方，階級値と度数の関係は，右の表のようになっているから，平均値は，

$$\overline{x} = \frac{60 \times 3 + 70 \times 4 + 80 \times 3}{10} = 70$$

のようにも求めることができる．

階級値	度数
60	3
70	4
80	3

中央値 (メジアン)

例えば，年収 600 万円の人が 1000 人いる町に年収 20 億円のスポーツ選手が 3 人引っ越してきたとする．この場合，この町の平均年収は，

$$\frac{600 \times 10^4 \times 1000 + 20 \times 10^8 \times 3}{1003} = 1196.41 \cdots \times 10^4 \text{ (円)}$$

となる．しかし，平均年収が約 1196 万円であるからといって，この町で平均年収が 1196 万円の人にあわせた政策をとっても誰も喜ばない．なぜなら，どの人の年収も 1196 万円から大きく離れているからである．

このように，平均は少数の大きな値あるいは小さな値 (これを**はずれ値**という) によって大きく動かされる場合があり，必ずしもデータの性格を表しているとは限らない．このような場合は**中央値**が便利である．中央値とは，n 個のデータを小さい順に並べたときちょうど中央に位置する値のことであるが，これをより具体的に述べると次のようになる．

[3] このとき，当然 $f_1 + f_2 + \cdots + f_k = n$ である．

データの分析