新指導要領　数学 I　データの分析
「四分位偏差」に関連した考察

吉田　一 *

2009 年 3 月 20 日

改訂指導要領の数学 I に統計の内容を教える「データの分析」が入りました．大項目は 2 つ,「ア　データの散らばり」と「イ　データの相関」です.「データの散らばり」のなかに記されている「四分位偏差」は指導要領には初登場の内容です．高校の数学教師でも耳慣れない用語かもしれません．そこで,「四分位偏差」に関連した説明と考察をします.

1　中央値と四分位値

中央値（median）は大きさの順に並んだデータを，真ん中の切れ目で半分ずつに 2 等分するところの値です．現行指導要領では，中学校の内容になく，高校でもほとんど履修されない選択科目の内容でしたが，新指導要領では最頻値などと共に中学校 1 年の「資料の活用」で扱われます．中央値や四分位値を考えるとき，データは大きさの順に並べ直す必要があります．ふつうは昇順（小さいほうから大きいほうへ）に並んでいるものとします.

四分位値（quartile）は大きさの順に並んだデータを 4 等分するところの値です．4 等分ですから，切れ目は 3 つあります．**下四分位値**（lower quartile）はデータ全体の小さいほうから $\dfrac{1}{4}$ に位置する値であり，**上四分位値**（upper quartile）はデータ全体の大きいほうから $\dfrac{1}{4}$ $\left(\text{小さいほうから}\dfrac{3}{4}\right)$ に位置する値です．また，下四分位値を第 1 四分位値，上四分位値を第 3 四分位値とも呼びます．中央値は第 2 四分位値ということになります．第 1 四分位値，第 2 四分位値，第 3 四分位値はそれぞれ Q_1，Q_2，Q_3 の略記号が使われます．指導要領には直接「四分位値」の記載はありませんが,「四分位偏差」を扱うなら当然欠かせない概念として取り上げなくてはなりません.

四分位値はさらに一般化して，$0 \leqq p \leqq 1$ である値 p に対しても用いられます．ふつう p はパーセントを用い，パーセンタイル値（percentile）と呼びます．たとえば，10 パーセンタイル値とは，小さいほうから 10 パーセント目に位置する値です．10 パーセンタイル値を境に，データは小さいほうから 10 %のデータ群と大きいほうから 90 %のデータ群に 2 分割されます．下四分位値は 25 パーセンタイル値，中央値は 50 パーセンタイル値，上四分位値は 75 パーセンタイル値です.

*YOSHIDA Hajime　河合塾講師　cq2h-ysd@asahi-net.or.jp

2　四分位範囲と四分位偏差

データの範囲（range）は 最大値（maximum value; Max）と最小値（minimum value; Min）の差，すなわち Max − Min で表される値です．範囲も新指導要領では中学校 1 年の「資料の活用」で扱われます．

四分位範囲（inter-quartile range; IQR）とは第 3 四分位値と第 1 四分位値の差，すなわち $\mathrm{Q}_3 - \mathrm{Q}_1$ で表される値です．両端の半分のデータを捨てた中央部半分のデータの範囲です．**四分位偏差**（quartile deviation）とは四分位範囲の半分，$\dfrac{\mathrm{Q}_3 - \mathrm{Q}_1}{2}$ をいいます．

四分位範囲や四分位偏差は，代表値を中央値としたときの，データのちらばりを示す値として使われます．分布の端のほうには極端に外れた値が存在することもあり，平均値や標準偏差はその外れた値に影響されますが，中央値や四分位範囲は外れた値の影響を受けにくい値です．また，分散や標準偏差に比べて値を求めるための計算量は少なくてすみます．[*1]

ただし，四分位偏差は中央値を中心とした指標とはいえません（言うまでもなく，平均値中心でもありません）.［平均値±標準偏差］と同じようなものとして［中央値±四分位偏差］とはとらえられません．分布が左右対称でない場合を考慮すれば，四分位偏差よりも四分位範囲を考えるほうがよいといえます．[*2]

3　四分位値の異なる解釈

中央値の場合，データの総数が奇数個の場合には「ちょうど真ん中」のデータが存在し，その値が中央値です．偶数個の場合には「ちょうど真ん中」のデータは存在しないので,「ちょうど真ん中」を挟む 2 つの値の平均値を中央値としました．これと同様の処理が四分位値でも必要になります．感覚的に言えば「四分位値は中央値で半分に分けたデータの中央値」なのですが，データが偶数個，つまり「ちょうど真ん中」がない場合には，解釈が分かれます．

第一の解釈は中央値のデータがある場合にはそれを含めて処理しますが，ない場合には中央値の情報は除き，実際に存在するデータだけで処理をします．第二の解釈は，中央値のデータがあるかないかに関わらず，その中央値の情報を使って処理をします．次のデータ例で説明します.「データのちょうど真ん中」の有無によって 4 通りのパターンがあります．

【A 1】データの総数が $4k+1$ 個の場合（k は自然数，以下も同じ）
（元のデータにも，二分割したデータにも「ちょうど真ん中」がある）

データ番号 n	1	2	3	4	5	6	7	8	9
データ x_n	3	4	6	8	9	9	10	12	15

中央値は 5 番目の値で，9 です．この場合は，どちらの解釈でも下四分位値は 1 番目から 5 番目の 5 個のデータの中央値，上四分位値は 5 番目から 9 番目の 5 個のデータの中央値となるので，$\mathrm{Q}_1 = x_3 = 6$, $\mathrm{Q}_3 = x_7 = 10$ です．

[*1] その代わり，データの並べ換えが必要です．
[*2] だから，指導要領も四分位偏差ではなく四分位範囲であってほしかったのですが．

【Ａ２】データの総数が $4k+3$ 個の場合
（元のデータには「ちょうど真ん中」があるが，二分割したデータにはない）

データ番号 n	1	2	3	4	5	6	7
データ x_n	3	4	6	8	9	10	13

中央値は 4 番目の値で，8 です．この場合も，どちらの解釈でも下四分位値は 1 番目から 4 番目の 4 個のデータの中央値，上四分位値は 4 番目から 7 番目の 4 個のデータの中央値となるので，$\mathrm{Q}_1 = \dfrac{x_2+x_3}{2} = \dfrac{4+6}{2} = 5$, $\mathrm{Q}_3 = \dfrac{x_5+x_6}{2} = \dfrac{9+10}{2} = 9.5$ です．

このように，データの総数が奇数個の場合には，問題はありません．

【Ｂ１】データの総数が $4k+2$ 個の場合
（元のデータには「ちょうど真ん中」がないが，二分割したデータにはある）

データ番号 n	1	2	3	4	5	6	7	8	9	10
データ x_n	3	4	6	8	9	10	11	12	14	17

中央値は 5 番目と 6 番目の間で，$\dfrac{x_5+x_6}{2} = \dfrac{9+10}{2} = 9.5$ です．

第 1 の解釈で四分位値を求めます．下四分位値は 1 番目から 5 番目までの 5 個のデータの中央値，上四分位値は 6 番目から 10 番目までの 5 個のデータの中央値となるので，$\mathrm{Q}_1 = x_3 = 6$, $\mathrm{Q}_3 = x_7 = 12$ です．

次に，第 2 の解釈で四分位値を求めます．この解釈では，中央値はデータの 5.5 番目の値とみるのです．つまり，データの位置（小さいほうから何番目か）を整数値だけをとる番号ではなく，連続的な目盛りであると考えます．下四分位値は 1 番目の値と 5.5 番目の値の中央，すなわち $\dfrac{1+5.5}{2} = 3.25$ 番目の値とし，3 番目のデータ $x_3 = 6$ と 4 番目のデータ $x_3 = 8$ の間を 0.25 : 0.75 に比例配分した 6.5 となります．上四分位値は 5.5 番目の値と 10 番目の値の中央，$\dfrac{5.5+10}{2} = 7.75$ 番目の値とし，7 番目のデータ $x_7 = 11$ と 8 番目のデータ $x_8 = 12$ の間を 0.75 : 0.25 に比例配分した 11.75 となります．すなわち，$\mathrm{Q}_1 = 6.5$, $\mathrm{Q}_3 = 11.75$ です．

【Ｂ２】データの総数が $4k$ 個の場合
（元のデータにも，二分割したデータにも「ちょうど真ん中」がない）

データ番号 n	1	2	3	4	5	6	7	8
データ x_n	3	5	6	8	9	9	11	14

中央値は 4 番目と 5 番目の間で，$\dfrac{x_4+x_5}{2} = \dfrac{8+9}{2} = 8.5$ です．

第 1 の解釈での下四分位値は 1 番目から 4 番目までの 4 個のデータの中央値，上四分位値は 5 番目から 8 番目までの 4 個のデータの中央値で，$\mathrm{Q}_1 = 5.5$, $\mathrm{Q}_3 = 10$ です．

第 2 の解釈では中央値はデータの 4.5 番目の値とみます．下四分位値は $\dfrac{1+4.5}{2} = 2.75$ 番目の値，上四分位値は $\dfrac{4.5+8}{2} = 6.25$ 番目の値とし，$\mathrm{Q}_1 = 5.75$, $\mathrm{Q}_3 = 9.5$ です．

このように，データが偶数個の場合には，解釈の違いにより四分位値の値に違いが起きてしまいます．

実は，第一の解釈による値は**ヒンジ値**（hinge）と呼ばれます．[*3]　すなわち，

下ヒンジ値：　中央値以下のデータの中央値
上ヒンジ値：　中央値以上のデータの中央値

です．第二の解釈は，連続的なパーセンタイル値にも拡張して使えるので，より一般的な考え方といえるでしょう．これは「内分」の考え方なのですが，数学Ⅱの「図形と方程式」で扱われることになっています．したがって，中学生，高校１年生に指導する場合には，実際に存在するデータで考えればよいヒンジ値のほうが適しているでしょう．補正の計算を必要とする四分位値の指導は難しいのではないでしょうか．かといって，データ処理に Excel を利用すると，そこで使われている quartile 関数や percentile 関数は第二の解釈での結果です．生徒にも教師にも混乱が起きないように注意が必要です．

	A	B	C	D	E	F	G	H	I	J	K
1											
2		No.	1	2	3	4	5	6	7	8	
3		データ	3	5	6	8	9	9	11	14	
4											
5		最小値	3		=MIN(C3:J3)						
6		下四分位値	5.75		=QUARTILE(C3:J3,1)						
7		中央値	8.5		=MEDIAN(C3:J3)						
8		上四分位値	9.5		=QUARTILE(C3:J3,3)						
9		最大値	14		=MAX(C3:J3)						
10			※上記の値はデータ範囲を固定して求められる.								
11		下ヒンジ値	5.5		=MEDIAN(C3:F3)						
12		上ヒンジ値	10		=MEDIAN(G3:J3)						
13			※ヒンジ値を求めるにはデータ範囲を指定し直す.								

図 1　Excel で四分位値を求める

でも，この違いは「どちらが正しいか」という問題ではありません．どちらでもたいした差はない，という感覚が必要です．[*4]　分布のようすをつかむことが大きな目的なのであり，四分位値を求めること自体が目的ではないからです．追って発行される『指導要領解説』ではどう記載されるのでしょうか．

付記：「データの分析」が履修されない可能性

せっかく統計分野が必修科目に入ったのだけれど，すべての高校生が学ぶとは限らないことを指摘しておきます．数学Ⅰの履修は３単位．でも，２単位に減らしてもよいことになっています．[*5]　２単位で３単位分の内容を学習することはないのですから，何かの内容が削られるとすると，「データの分析」はその第一候補となりかねません．[*6]

[*3] ヒンジとは蝶番（ちょうつがい）．この位値でデータを折り返して畳む感じから名付けられたそうです．
[*4] 平均点や偏差値のちょっとした違いを気にする教師には大問題かもしれませんが．
[*5] 指導要領 総則 第３款に記載．国語，英語の必修科目も同じく単位減が可能になっています．
[*6] 「2 単位の時間で 3 単位の内容を学習する」という解釈もあるようです．文部省の意図はどちらでしょうか?