数学Ⅰ「データの分析」の教材について

埼玉県立大宮武蔵野高等学校　太田　敏之

＜要旨＞

来年度から新学習指導要領が先行実施され、数学Ⅰで「データの分析」について新しく学習する。そこで本論では、身近な話題を取り入れて考察することから、生徒が「データの分析」に興味をもち、有用性を感じるような教材について研究して提案する。

１．はじめに

高等学校の数学については、来年度から新学習指導要領が先行実施され、数学Ⅰで「データの分析」について新しく学習する。しかし、近年高等学校で統計分野を扱ってきた場面は少なく、統計分野についての教材研究をする必要がある。また新学習指導要領では、課題学習や言語活動を取り入れることについても考慮する必要がある。

そこで本論では、身近な話題を取り入れて考察することから、生徒が「データの分析」に興味をもち、有用性を感じるような教材について研究して提案することとする。

２．「データの分析」の指導法

「データの分析」の単元の内容は大きく分けて、「データの散らばり」と「データの相関」がある。

「データの散らばり」では、中学校で学習した内容の復習として「度数分布表、ヒストグラム、相対度数、平均値、中央値（メジアン）、最頻値（モード）」といった内容と、高等学校で新たに学習する「四分位数、箱ひげ図、範囲（レンジ）、分散、標準偏差」といった内容を学習する。

また、「データの相関」では、「散布図、共分散、相関係数」といった内容を学習する。

身近なデータを分析することが多いので、生徒が興味をもちやすい単元ではあるが、単に数値計算で終わるのではなく、データを分析してわかることを考察したり議論したりするなどの言語活動を効果的に取り入れて、生徒がデータの分析の有用性を感じるような教材や指導法の工夫が必要であると考える。

また、実際のデータを分析すると、数値計算が煩雑になる場合があるので、場面によって効果的にコンピュータや電卓などの機器を取り入れる必要があると考える。

３．「データの分析」の教材例

ここでは、「データの分析」についてのいくつかの教材について提案する。

＜教材例１＞平均点

＜平均点を取ると、順位は真ん中か？＞

クラスでＡ君はテストで平均点ちょうどだった。Ａ君はそのことを家で報告すると、
「じゃあ、ちょうどクラスで真ん中くらいの順位だね。」と言われた。

果たして本当にそう言い切れるだろうか？

この教材では、生徒の班を作り、答えを予想させ、どのようなケースがあるかを議論させ、そう言い切れないという考えの班には、そうなるような問題例を作成させ発表させるとよい。

例えば、40 人のクラスで、**平均点が 50 点、Ａ君も 50 点**だとする。極端でわかりやすい例を考えると、55 点の生徒が 34 人いて、Ａ君が 50 点で、20 点の生徒が 3 人いて、10 点の生徒が 2 人いるとすると、平均点は(55×34＋50＋20×3＋10×2)/40＝50(点)となる。しかし、Ａ君の順位は **35 位**である。

これがもし大学の入学試験の場合は、平均点を取っても不合格になってしまう場合がありえるのである。このようなケースの問題作りを通して、平均の理解を深めるとよい。

そこで、上記のような例の場合、平均値ではなく、中央の順位にある数値である、**中央値(メジアン)**を求めることが大切になってくる。

上のデータを少し変えて、次のようなデータを考えてみる。

得点	**58**	56	**54**	52	**50**	48	20	10
人数	13	6	**2**	12	**1**	1	3	2

平均値は(58×13＋56×6＋54×2＋52×12＋50＋48＋20×3＋10×2)/40＝50(点)となる。Ａ君の点数は **50 点**で平均値と同じだが、順位は **34 位**と低くなる。このとき、**中央値(メジアン)**は 40 人の真ん中の 20 番目と 21 番目の点数の平均の **54 点**である。

よって、点数が中央値(メジアン)と同じ **54 点**の人は、順位はちょうど真ん中であることがわかる。

つまり平均値ではなく、中央値(メジアン)の 54 点より上かどうかで、自分の位置を考えるとよいことがわかる。

また、度数がもっとも多い階級の階級値である、**最頻値(モード)**についても考察するとよい。このデータでは、最頻値(モード)は、度数が一番多い **58 点**となる。58 点はクラス 1 位の点数であるが、58 点を取った人が一番多いので、もっとも多数派で、一般的な点数であることがわかる。

＜教材例２＞平均収入

＜平均年収は意味をもたないことも多い？＞
ある町に住む 40 歳 100 人の平均年収は 1600 万円で、Ｂさんの年収は 800 万円であるとする。
Ｂさんの年収は、平均年収の半分なのだが、Ｂさんの年収は低いと言い切れるだろうか？

この教材でも、生徒の班を作り、答えを予想させ、どのようなケースがあるかを議論させ、そう言い切れないという考えの班には、そうなるような問題例を作成させ発表させるとよい。

例えば、100 人の年収が以下のような分布だったとする。

年収の階級値	200 万	400 万	600 万	800 万	1000 万	…	10 億
人数	11	23	26	30	9		1

平均値は、(200×11＋400×23＋600×26＋800×30＋1000×9＋100000)/100＝**1600(万円)**となる。

この平均値とＢさんの年収 800 万円を比べると、確かにＢさんの年収は低いように感じるが、このデータにおけるこの平均値はあまり意味をもたないと考えられる。

そこで、**中央値(メジアン)**を求めると、中央値(メジアン)は、階級値が **600 万円の階級**にあり、中央値とＢさんの年収 800 万円を比べると、Ｂさんの年収の方が高く、Ｂさんは真ん中より上の順位にいるということがわかる。

つまり、データの分布が正規分布に近い分布であれば、平均値は大きな意味をもつが、教材例１の例ように、分布が正規分布から大きく外れていたり、この例のように、**年収 10 億円**という野球のメジャーリーガー級の給料をもらう人（このようなデータを「外れ値」という）が一人でもいたりすれば、平均値はあまり意味をもたないことがわかる。

また、**最頻値(モード)**を求めると、この場合の最頻値(モード)は 30 人いる**年収 800 万円**である。 つまり、最頻値（モード）の年収と同じであるＢさんは多数派であるということになり、一般的な年収の人であるということがいえるのである。

このように、データの分布や内容、何について分析するかによって、平均値と中央値(メジアン)、最頻値(モード)を使い分ける必要があることがわかる。

作られた都合のよいデータではなく、より身近なデータを使うと、生徒はより興味をもち、有用性を感じると考える。

例えば、プロ野球「読売ジャイアンツ」の選手89 人（育成選手含む）の平成 23 年の推定年俸のデータがある。1 位の選手の推定年俸は 4 億 5000 万円、2 位 4 億 3000 万円、3 位 4 億・・・であり、76 位に 240 万円が 14 人いて、これが最低年俸である。**平均推定年俸(平均値)**を計算すると **4359 万円**であるが、4300 万円以上もらっている選手は 20 人しかおらず、**中央値(メジアン)**の 45 位の選手の推定年俸は **800 万円**である。

「プロ野球の選手の平均推定年俸 4359 万円」というデータからは、プロ野球選手はとてもお金が稼げる職業のように思えるが、「中央値が年俸 800 万円」となると、一流選手にならなければそれほどお金が稼げるとはいえないことが考察できる。

また、**最頻値(モード)**は、14 人いる **240 万円**で、これは育成選手であり最低年俸の選手たちである。つまり、このような年俸の低い厳しい契約の選手たちがもっとも多いことがわかる。

このような現実のデータを扱うと、より有用性を感じることができるが、計算の量が多くなるので、コンピュータや電卓などの機器を利用するとよい。

＜教材例３＞バスケットボールの選手比較

＜どの選手と契約する？＞
次のデータは、A，B，Cの３選手の、最近 10 試合のバスケットボールの試合で挙げた点数を低い順に並べたものである。

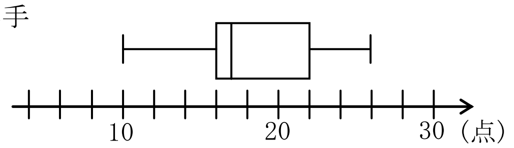
A選手	10	14	16	16	16	18	18	22	24	26
B選手	4	6	10	14	14	20	24	28	30	30
C選手	12	14	14	16	16	20	20	22	22	24

これらのデータについて、次の値をそれぞれ求め、箱ひげ図を書き、３選手の特徴について考察せよ。
①平均値
②第２四分位数
③第１四分位数
④第３四分位数
⑤範囲（レンジ）
⑥四分位範囲と四分位偏差
⑦分散

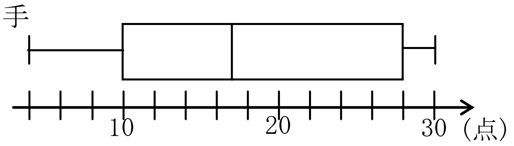
この教材でも、各生徒でそれぞれの値を求め、箱ひげ図をかいた後、生徒の班を作り、３選手の特徴について議論させるとよい。

①平均値は、３選手とも 18 点で同じである。
②第２四分位数は、データ全体の中央値なので、
　Ａ選手；(16＋18)/2＝17（点）
　Ｂ選手；(14＋20)/2＝17（点）
　Ｃ選手；(16＋20)/2＝18（点）
より、Ｃ選手が一番高い。
③第１四分位数は、前５つのデータの中央値なので、
　Ａ選手;16 点，Ｂ選手;10 点，Ｃ選手;14 点
より、Ａ選手が一番高い。
④第３四分位数は、後５つのデータの中央値なので、
　Ａ選手;22 点，Ｂ選手;28 点，Ｃ選手;22 点
より、Ｂ選手が一番高い。
⑤範囲は(最大値)－(最小値)なので、
　Ａ選手;26－10＝16(点)
　Ｂ選手;30－ 4＝26(点)
　Ｃ選手;24－12＝12(点)
より、Ｂ選手が一番大きく、Ｃ選手が一番小さい。
⑥四分位範囲は(第３四分位数)－(第１四分位数)、四分位偏差は(四分位範囲)÷2 なので、
　Ａ選手;22－16＝ 6(点)，6÷2＝3(点)
　Ｂ選手;28－10＝18(点)，18÷2＝9(点)
　Ｃ選手;22－14＝ 8(点)，8÷2＝4(点)
より、Ｂ選手が一番大きい。
⑦分散をそれぞれ計算すると、
Ａ選手;20.8(点)，Ｂ選手;86.4(点)，Ｃ選手;15.2(点)
より、Ｂ選手が一番大きく、Ｃ選手が一番小さい。
◎箱ひげ図

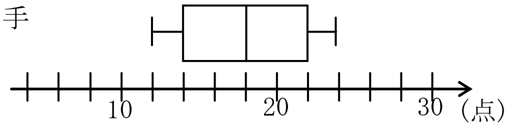
Ａ選手



Ｂ選手



Ｃ選手



以上の値をもとにデータを分析すると、３選手の得点の**平均値**は同じであるが、Ｃ選手は**中央値**が高く、**範囲**や**分散**が小さいため、安定した選手であると考えることができる。また、Ｂ選手は**第３四分位数**が大きく、**範囲**や**分散**が大きいため、調子に波があるが、調子がいいときはたくさん点をとれる爆発力がある選手であると考えることができる。Ａ選手は、Ｂ選手とＣ選手の中間であり、**第１四分位数**が高いことから、調子が少し悪いときが一番少ない選手と考えることができる。

また、箱ひげ図を見ると、データをより視覚的に分析することができる。

このデータの分析にはいろいろな見方があり、これが正解というものはないが、いろいろと議論してデータを分析するとよい。また、これらのデータをもとに、あなたが監督ならばどの選手と契約するかとか、どの選手をどのような場面で起用するかなどを具体的に議論してみてもよい。

＜教材例４＞アイスクリームとおでんの売上

下の表はある年の８月の 10 日間の最高気温とある店のアイスクリームの売り上げの表である。このデータの最高気温（x℃）とアイスクリームの売り上げ（y 万円）の相関係数 r を、小数第 3 位を四捨五入して答えよ。

x℃	30	32	33	35	36	34	32	31	33	34
y 万円	19	24	25	29	31	26	23	21	24	28

平均値：$\bar{x} = \dfrac{30+32+\cdots+33+34}{10} = \dfrac{330}{10} = 33$

平均値：$\bar{y} = \dfrac{19+24+\cdots+24+28}{10} = \dfrac{250}{10} = 25$

x の標準偏差 s_x，y の標準偏差 s_y は、

$$s_x = \sqrt{\frac{1}{10}\left\{(30-33)^2+(32-33)^2+\cdots+(34-33)^2\right\}} = \sqrt{3}$$

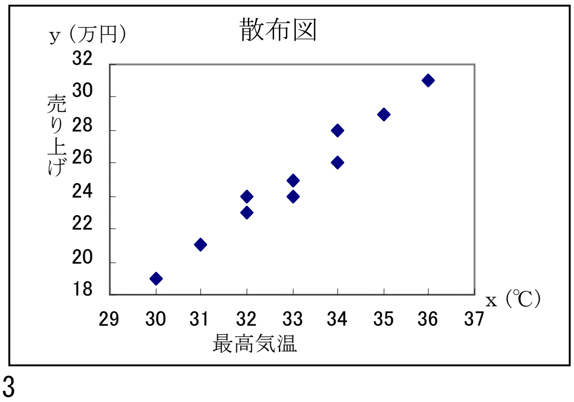
$$s_y = \sqrt{\frac{1}{10}\left\{(19-25)^2+(24-25)^2+\cdots+(28-25)^2\right\}} = 2\sqrt{3}$$

x, y の共分散：$s_{xy} = \dfrac{1}{10}\{(30-33)(19-25)+$

$$+(32-33)(24-25)+\cdots+(34-33)(28-25)\} = \frac{59}{10}$$

よって相関係数は、

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\frac{59}{10}}{\sqrt{3}\times 2\sqrt{3}} = \frac{59}{60} = 0.983\cdots \fallingdotseq 0.98$$



相関係数が１に近いので、**正の相関**である。つまり、「気温が高ければ高いほどアイスクリームは売れる」という分析をすることができる。

このデータについて散布図をかくと、正の相関であることが視覚的にわかる。

下の表はある年の１月の 6 日間の最高気温とある店のおでんの売り上げの表である。このデータの最高気温（x℃）とおでんの売り上げ（y 万円）の相関係数 r を、小数第３位を四捨五入して答えよ。

x℃	10	7	6	12	9	10
y 万円	13	19	19	12	15	12

平均値：$\bar{x}=\dfrac{10+7+6+12+9+10}{6}=\dfrac{54}{6}=9$

平均値：$\bar{y}=\dfrac{13+19+19+12+15+12}{6}=\dfrac{90}{6}=15$

x の標準偏差 s_x， y の標準偏差 s_y は、

$$s_x=\sqrt{\frac{1}{6}\{(10-9)^2+(7-9)^2+\cdots+(10-9)^2\}}=\sqrt{4}=2$$

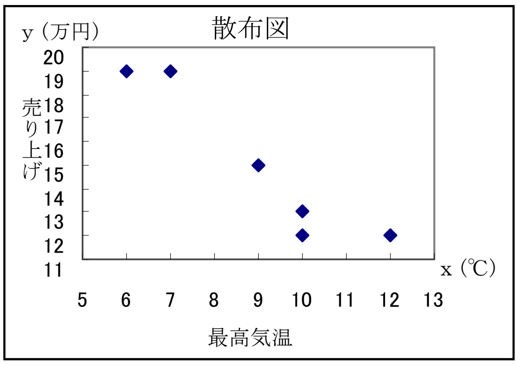
$$s_y=\sqrt{\frac{1}{6}\{(13-15)^2+(19-15)^2+\cdots+(12-15)^2\}}=\sqrt{9}=3$$

x, y の共分散：$s_{xy}=\dfrac{1}{6}\{(10-9)(13-15)+(7-9)(19-15)+$

$$\cdots+(10-9)(12-15)\}=-\frac{17}{3}$$

よって、相関係数は、

$$r=\frac{s_{xy}}{s_x s_y}=\frac{-\frac{17}{3}}{2\times 3}=-\frac{17}{18}=-0.944\cdots\fallingdotseq -0.94$$



相関係数が今度は－１に近いので、**負の相関**である。つまり、「気温が低ければ低いほどおでんは売れる」という分析をすることができる。

このデータについても散布図をかくと、負の相関であることが視覚的にわかる。

４．まとめと今後の課題

「データの分析」の単元は、他の単元に比べて身近な内容を扱いやすい単元なので、教材をうまく用意し、扱い方を工夫することで、生徒が数学に興味をもち学習意欲を高めやすく、数学の有用性を感じやすい単元であると考える。

また、データを分析してわかることを考察したり議論したりするなどの活動を通して、生徒が数学の有用性を感じるとともに、新学習指導要領で挙げられている言語活動を効果的に取り入れることができると考える。

しかし、「データの分析」で扱う教材は、その内容や計算量などの観点で、各学校の生徒にあった教材やデータを用意するのが簡単ではないと考える。

例えば、現実のデータやそれに近いデータを用意すると、データの量が多くなり、コンピュータや電卓などの機器を利用しないと分析するのが大変で、その機器の用意など、準備も大変である。一方、机上での演習や定期考査など機器を利用しないで解かせるための問題は、データ量が少なく、計算がある程度簡単になる必要があるが、そのような問題は現実のデータを加工したり、作成したりするのが大変である。また、データ量が少ないために分析がきちんとできなかったり、現実離れした問題になってしまったりすることも多く、注意が必要である。

よって今後の課題は、来年度から始まるデータの分析の授業実践を通じて、教材や扱うデータを精選したり開発したりしていくとともに、多くの先生方と情報交換をして、教材を数多く共有していくことであると考える。よって、研究会やホームページなどを通して、多くの教材を提案していきたいと思っている。

＜参考・引用文献＞

[1]東京書籍編集部(2011),「ニューサポート数学Ⅰ＋Ａ」.

[2]太田敏之 ,「高校数学教材集」,
http://www7b.biglobe.ne.jp/~math-tota

[3]「プロ野球データ Freak 読売ジャイアンツ年俸ランキング」
http://baseball-data.com/ranking-salary/g/