データ分析における課題学習の研究

愛媛県立今治工業高等学校　藥眞寺裕

１．はじめに

昨年度より数学Ⅰに新たに「データの分析」が導入された。旧課程の「数学B」（統計とコンピュータ）の内容がほぼ移行したものである。しかしながら、旧課程では他の内容を選択する学校が多く、授業で統計分野を扱ってきた場面は少ない。昨年度は試行錯誤をしながら授業を行った先生も私を含めて多いと思う。そのため、統計分野についての教材研究をすることが急務である。また、課題学習については,「それぞれの内容との関連を踏まえ、学習効果を高めるよう適切な時期や場面に実施するとともに、実施に当たっては数学的活動を一層重視するものとする。」とあり、より「数学のよさ」を生徒に感じさせることができることについても考慮する必要がある。

そこで本研究では、身近な話題を取り入れ考察することにより、生徒が「データの分析」に興味を持ち、有用性を感じることができることを目的とする。また、本校は工業高校であり、専門科目で「数学」を使う場面は多くあるが、「数学」を苦手とする生徒が多い。そのため、生徒の計算力を考慮すると、短時間で大きなデータや多量のデータを扱うことができ、数学的な活動における試行錯誤が生徒の負担にならず、生徒が自ら考える機会が増える授業展開を考える必要性がある。そして、限られた授業時数の中で、課題学習を行うためには、授業１時間でまとめることのできるシンプルな内容でもいいのではと考えた。以下の教材例では、生徒に工業科の科目で使用している電卓を持ってくるように指示し、必要であれば使ってもよいと指導していることを仮定している。

２．統計にだまされないように

統計学とは何か？という質問を受けた時、「たくさんあるデータの中から、有用な情報を取り出す方法」と答えるようにしている。いくらたくさんのデータがあっても、データそのものがあるだけでは情報にはなりえない。統計学的手法によって処理されて初めて「人の行動の判断材料」である「情報」になる。現代の社会生活を営む上で、怪しげな情報にだまされないためにも統計の知識は必要である。そのことを生徒に体験・実感する教材の研究が必要である。以下の身近な「統計的表現」の意味するところや真偽について考察しても面白いかもしれない。いくつかの事象については、このあとの課題学習の教材例として、深く考察することにする。

【身近な統計的表現】

①「明日の降水確率は９０％。明日は雨降るね。」
②「日本人の平均寿命は８０歳」
③「４０人で数学の試験をして、私の得点は平均点と同じであった。私は、クラスで２０位である。」
④「あるプロサッカーチームの平均年俸が１０００万です。A選手の年俸が、７００万です。A選手の年俸は低い？」

３．平均点と中央値を用いた教材例

先日、２年生の数学Ⅱの第２学期中間考査が行われた。昨年度、数学Ⅰの授業で「データの分析」を学習した生徒たちである。私は、ある２つの学科の生徒の授業を担当しており、試験範囲を合わせ、同一問題で実施している。採点を進めている中で、面白いデータが得られたと思い、以下のような授業展開を考えた。

【数学Ⅱの第２学期中間考査の結果】

クラスA：受験者３６名、８０点以上１１名、平均点６８点
クラスB：受験者３８名、８０点以上１９名、平均点６５点

この結果をクラスBの担任の先生に説明し、SHRでクラスの生徒に話してもいいかと言われたので、それを了解し、その時の生徒の反応を教えてくださいとお願いした。数学を得意としている生徒は、高得点者がたくさんいるのに平均点が上がらない状況はなぜなのかということを理解していたが、そうでない生徒や意味を深く理解できなかった生徒には、高得点者が多数いてすごいと感じているだけの生徒もいたそうであった。そこで、以下のテーマを考えた。

【テーマ１：平均点を取ると、順位は真ん中か？】

クラスBでP君はテストで平均点ちょうどだった。P君はそのことを家で報告すると、「じゃあ、順位はちょうどクラスで真ん中だね。」と言われた。果たして本当にそのように言い切ってもよいのだろうか？

このテーマでは、様々なアプローチの仕方があり、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場合もある。まず、クラス別に中間考査の得点のヒストグラムをさせる。個々のデータの表示は、ここでは省略することにし、ヒストグラムは【図１】のようになる。

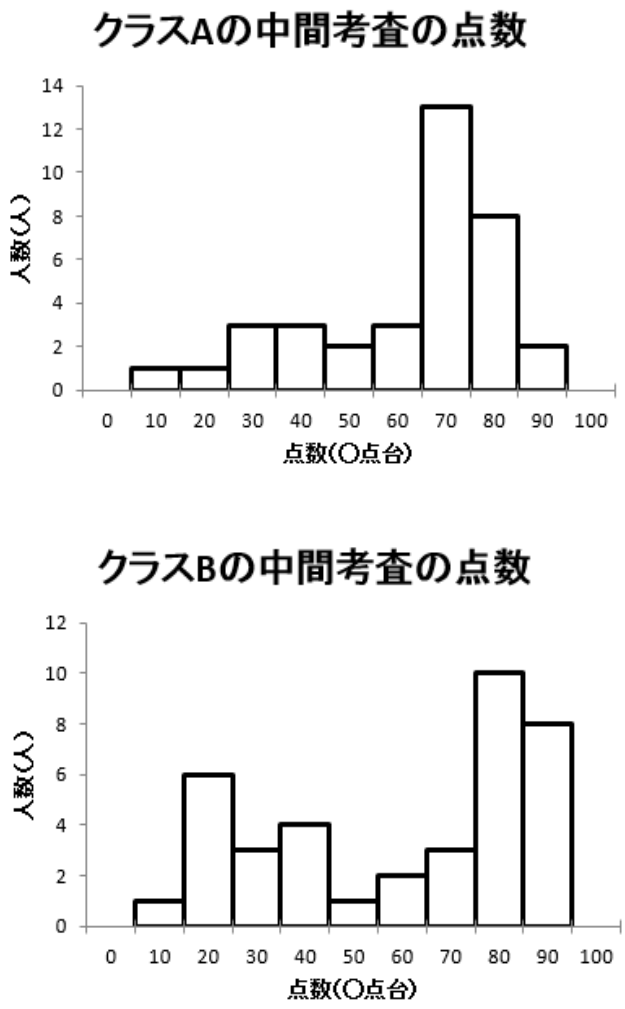
次に平均値と中央値（メジアン）を計算する。データ数が４０近くになるので、ここでは電卓使用を想定している。２つのクラスの平均値と中央値は以下の表のようになる。

クラス	平均点	中央値
A	６８．２	７６．０
B	６５．１	７８．０

P君の点数は６５点で平均値と同じだが、順位は２３位と低く

なる。このとき、中央値(メジアン)は３８人の真ん中の１９番目と２０番目の点数の平均の７８点である。よって、点数が中央値(メジアン)と同じ７８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７８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。

【図１：中間考査の得点のヒストグラム】



そして、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こともできる。このデータでは、最頻値(モード)は、度数が一番多いクラス A では７０点台、クラス B では８０点台となる。それぞれ最頻値(モード)は、平均点とかけ離れている点数であるが、得点者が一番多いので、もっとも多数派で、一般的な点数であることが言える。

また、習熟度のやや高い集団であれば、例えば、このテーマに対し、「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。」というアプローチもできると考える。以下のデータを想定し、

（１問１０点、１０問のテストを４０人が受験）

得点	10	20	30	40	50	60	70	80	90
人数	5	1	1	0	0	1	32	0	0

４０人のクラスで、P 君は６０点だとする。平均点を計算すると、

$$\bar{x} = \frac{10 \times 5 + 20 \times 1 + 30 \times 1 + 60 \times 1 + 70 \times 32}{40} \times 60$$

となるが、P 君の順位は３３位である。もしこれが入学試験の場合は、P 君は不合格になってしまう場合があり得る。（無論、このような試験が実際に行われた場合は、その試験の作成過程で問題があると判断される。）このようなテーマの問題作りを通して、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)や最頻値(モード)を求めることが大切であることを伝え、平均の理解を深めることができる。また、最頻値(モード)は、ヒストグラムが双峰性のときは、あまり意味をなさない代表値であることが分かる。

【テーマ２：平均収入は意味を持たないことがある？】

ある団地の各住人５０人の平均月収は４０万円で、A さんの月収は２０万円であるとする。A さんの月収は、平均月収の半分なのだが、A さんの月収は低いと言い切れるだろうか?

このテーマでも、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展開が考えられる。平均の欠点として、「突出した値に左右されやすい」という欠点がある。例えば、５０人の月収が以下のような分布だったとする。

月収	10	15	20	25	30	35	40		1000
人数	4	4	28	10	3	0	0		1

平均点を計算すると

$$\bar{x} = \frac{10 \times 4 + 15 \times 4 + 20 \times 28 + 25 \times 10 + 30 \times 3 + 1000 \times 1}{50} \times 40$$

となる。この平均値と A さんの月収２０万円を比べると、確かに A さんの月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。このデータのように、月収１０００万円という高額の給料をもらう人（これを 「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理解できる。そこで、中央値を求めると、中央値は

階級値が ２０万円の階級にあり、中央値とＡさんの年収２０万円を比べると、大きな差は認められず、一般的な月収の人であるということがいえる。また、最頻値を求めると、この場合の最頻値は２８人いる月収２０万円である。つまり、最頻値の月収と同じであるＡさんは多数派であるということになり、一般的な月収の人であるということがいえる。

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、最頻値を使い分ける必要がある。また、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、テーマ１のように、分布が正規分布から大きく外れていれば、別の視点からデータの分析をしなければならない。

４．まとめと今後の課題

データの分析の単元では、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。また、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。また、データを分析して分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。

また、テーマ１で用いた中間考査のデータについて、別の視点から分析することもできる。少し専門的にはなるが、平均も分散も異なる２つの分布を比較して、「どちらがよりばらついているか」を表現する方法として、「変動係数」という考え方がある。

変動係数はk

$$k=\frac{\text{標準偏差}}{\text{平均}}=\frac{\sqrt{\text{分散}}}{\text{平均}}$$

で求めることができる。平均と分散を計算すると、

クラス	平均点	分散
A	６８．２	３５７．４
B	６５．１	７０３．１

となり、各クラスの変動係数は

$$k_A=\frac{\sqrt{357.4}}{68.2}=0.277$$

$$k_B=\frac{\sqrt{703.1}}{65.1}=0.401$$

となる。この計算結果を見ると、クラス内のバラツキに差があると判定できる（変動係数の差が０.１以内であれば、「分布に特異的な差は見られない」と一般的に判断される）。

このように代表値だけでなく、度数分布について「同じ平均・同じ標準偏差」に換算して、比較することもできる。現在の統計学は、いろいろな分析手段・方法が確立されて、大数の法則や中心極限定理で証明されている（証明は煩雑なので省略）。１つのデータ集団に対し、それらをどのように利用するかが大切になってくる。生徒により身近なデータ分析としては、このデータを用いて偏差値を求める課題学習も考えられる。あるデータからその分布の平均を差し引き、標準偏差で割って新しいデータを作ると、その平均は０、標準偏差は１となる。このように変換したデータを標準得点という。この標準得点を１０倍して５０を足し、平均５０点、標準偏差１０点に換算したものが偏差値である。これは、学力テストが１００点満点で行われることが多いため、３０～７０点あたりのなじみのある値で分布中の位置を表現するために考案されたものである。

今年度は、授業の進度の関係で、授業実践がまだできていないが、３学期に行うことになる「データ分析」の授業で課題学習を実践したいと考えている。「データの分析」で扱う教材は、その内容や計算量などを考慮すると、生徒の実態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られる。昨年度から感じている課題として、授業での演習や定期考査など計算機器を利用しないで解かせることを仮定している問題は、データ量が少なく、計算がある程度簡単にしないといけない必要性が生じる。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正しくできなく、現実離れした結果を生むことも多々ある。それに対し、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になる。また、パソコン教室などで授業をするとなると、準備も含めて様々な課題が生じると思われる。今後は、試行錯誤をしながらではあるが、「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、より「数学のよさ」を生徒に感じさせることができる課題学習の在り方を模索していきたい。