「データの分析」のセンター対策から

有朋高校単位制課程　大谷 健介

0 はじめに

データの分析については以前から何度かレポートを出させていただいています。その中で、「四分位数と分散を同列で取り扱うことはしない」「四分位数や箱ひげ図は、この世に出てから10数年しか経っていないため、その定義は世界に13もあって標準が存在しない」等の問題点についても、取り上げてきました。

そして、この1月にセンター試験の数学①で「データの分析」の分野としてはじめて出題されました。センター試験に至るまでに何人かの生徒とセンター対策をするにあたって、たくさんの問題に取り組んできました。きょうは、その過程で感じたことをレポートしたいと思います。

1 四分位数は出題されるのだろうか…

教科書に記されている四分位数は一律に同じ定義にしたがっている(文科省がこれって決めたやつ)。だから、受験生は「四分位数を求めなさい」と言われればみんな同じように考え同じように解答を導く。しかし、まともに四分位数を問うものが出た場合、定義(統計ソフト)によって答えが変わることが考えられ、その場合、社会的に説明がつかないと考えていました。だから、「四分位数を問うものは出題されないと思う」と生徒には(自信なさそうに)宣言していました。

したがって、積極的に取り組むは、分散と標準偏差と共分散、ときどき散布図に相関係数あたりです。

2 いろいろな対策問題に触れてみて

センターに特化してたくさんの問題に触れました。その中には出版社が時間配分や難易度にとても苦労しているあとが見られました。そのあたりをつまんでみます。

(1) データが5つなのに…

問題　次の表は、ある店で二つの商品A,Bの5日間の売り上げ個数をまとめたものである。データは整数値をとる。

日	1	2	3	4	5
商品A	12	11	14	20	13
商品B	11	P	13	14	Q

(1) 商品Aの平均値、分散、四分位範囲

(2) 商品Bの平均値が12、分散が2のときの中央値はQである。このときのP,Qの値

(3) AとBの相関係数に最も近いものを⓪～⑤から選ぶ

5つのデータに対しての(1)の問いはどうなのかなーと感じてしまいます。四分位範囲とは「中央値付近の50%のデータがどれくらいの散らばりか」を数値として表現するものなので、中央値付近の2.5個の散らばりを知るための数値と言うことになってしまします。「データの分析」の出題の難しさを感じる1問です。

(2) この問いは統計学として許されるのか…

問題　あるクラスの生徒5人が英語と数学の試験を受けた。いずれの試験も得点は負でない整数とし、満点は100点である。英語の試験の得点を変量x、数学の試験の得点を変量yとする。また、それぞれの平均値を$\bar{x}$,$\bar{y}$とする。次の表はこれらのデータをまとめたものである。ただし、一部のデータが失われており、空欄となっている。

番号	x	y	$x-\bar{x}$	$y-\bar{y}$
1	85			30
2	72	67		B
3	A			−20
4	30			−20
5	61		3	0
合計				0

(1) 変量xの平均値とAの値

(2) Bの値と変量yの分散

(3) xとyの散布図の選択

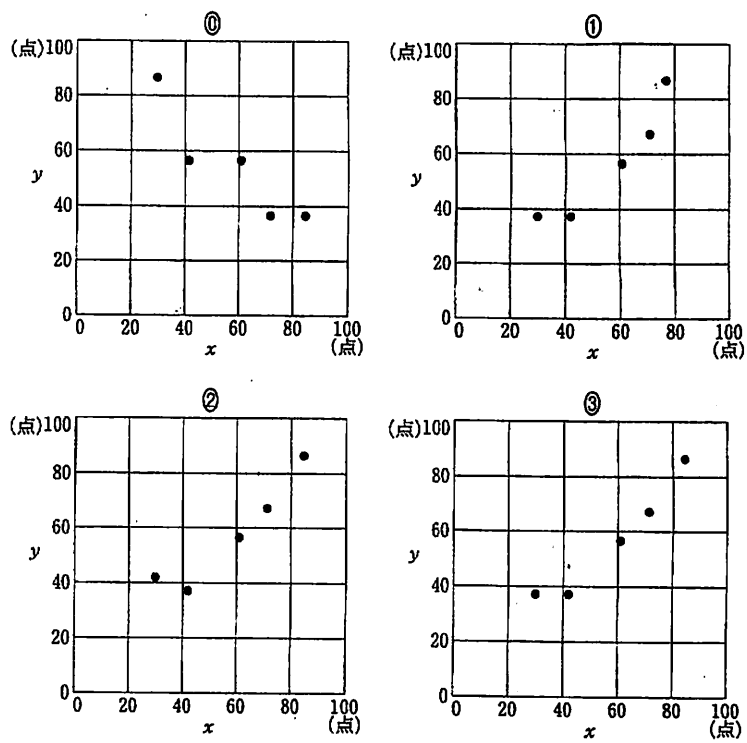
(4) xとyの相関係数の選択

⓪　−0.97

①　−0.04

②　0.04

③　0.97



これも5つのデータに対しての統計です。5つずつのデータに対して当てはまる散布図を選択するというのはやや寂しい問題ですが、いろいろな問題集に5〜6個程度からなる散布図の問題が散見されます。この程度であれば、単純に数字を比較しただけでも相関が読み取れそうなものです。

ところで統計を取るための大切なデータを失ってしまうとは…あとどこか1つでも失われていたら使えない資料になってしまうではないか…と言うのは考えすぎでしょうか。

(3) 計算ミスは許されない

問題　ある高校の生徒 10 人がテスト P、テスト Q を受けた。得点は負でない整数とし、10 人のテスト P、Q の得点をそれぞれ変量 x、y とする。次の表はこれをまとめたものである。

(1) A、B、C および変量 x の標準偏差

(2) 変量 y の第 1 四分位数と第 3 四分位数

(3) 変量 x と変量 y の相関係数

番号	x	y
1	61	60
2	71	62
3	75	75
4	66	59
5	59	63
6	69	64
7	65	58
8	61	77
9	63	56
10	60	56
平均値	A	63
中央値	B	61
分散	C	49

次はよく見られる傾向の対策問題です。データが増える分、少し見栄えが良くなる感じがします。しかし、先述のとおり、(2)のような出題はあり得ないと私は考えていました。

生徒が間違えたのは(3)です。$(x-\bar{x})(y-\bar{y})$ の計算ミスをしました。しかも 3 人いて 3 人とも 1 カ所だけ計算ミスをしました。単純な 20 個のひきざんと 10 個のかけざんの繰り返しで、そんなので間違ってはいけないのは百も承知ですが…何とも残念な気持ちになります。

ちなみにこの問題は分散や平均値、そして相関係数の数値がすっきりしていて扱いやすく、私は好きです。

(4) その他

そのほかにもいろいろな問題に触れましたが、統計検定からの対策問題では散布図が難しくて答えを見ても正解が理解できないという問題や、7 つのデータから第 1 四分位数を求めよと言う問題に驚くなど、新しい経験をしました。

3 センター試験の問題に思う

そして、センター試験です。はじめに問題を目にして、「げっ!!4 ページもある！こんなに出たのか」と、少し驚いたのですが、問題数がたくさんあるわけではなく(箱ひげ図や散布図にスペースがとられた)、配点も 15 点。私の予想はややはずれたものの、やはり四分位数では、どの定義でも同じ答えになるように工夫がなされ、度数分布表からの幅を持たせた出題となっていました。箱ひげ図についても同様で、(教科書の定義通りの)正しいものを選択するのではなく、「矛盾するもの」を選択させ、社会的に説明のつくような出題になっていたように感じます。また、先に指摘した「1 つの計算ミスも許されない」懸念は、統計理解の本質ではないことをクリアするために、いろいろな値を見せておいて、「この中から必要なデータを使って求めてね」という好ましい姿勢が伺えました。

ただ、やはり他の数学①の問題と比較するとすごく違和感を覚えます。この単元についてそこそこセンター対策しておくと容易に解ける問題でしたし、なによりすべてが"選択問題"でした。

今年度、私が触れたセンター対策問題集・模擬試験のほとんどは、これまでの数学Bの統計の問題をよりどころにしていたように感じましたので(K林館は統計検定からの編集)、1度実施されたことにより、どのように変化していくのかは楽しみなところです。

4 蛇足

数年前、数学Ⅰ・A の平均点が 51 点だったことがあります。試験会場では、終了の声とともに異様な雰囲気となり泣き出す受験生もいたそうです。ネット上でもたくさんの嘆きが聞かれ、あらら～と思っていました。

今回、数学Ⅱ・B でも同じようなことが起こったようです(平均点が 40 点を割る見込み)。ネット上でも同様で、各種ニュースでも話題になっていました。

$\sin 7\theta$ や S_{4m} のあたりは受験生にとっては、とてもいやだっただろうと察しますし、平均変化率や微分係数の出題にも面食らったに違いありません。

私もセンター対策をするにあたって、(余裕がないので)センターに特化して指導してしまいますが、それではうまくいかないことを改めて思い知らされました。もっと広く実力をつけてあげなければいけませんでした。

世間で言うところの「今年の数学○○の問題は…」というのは、詰まるところ、過去の傾向に沿っていれば「易しかった」、過去の傾向から大きく変化すると「難しかった」となるのだと思いました。60 分の限られた時間でスピードが要求される中、触れたことのない問題が来ると厳しい…と、それに対応する力をつけさせてあげなければならないと、今年の問題を見て大反省しているところです。

第3問 (必答問題) (配点 15)

〔1〕 ある高校3年生1クラスの生徒40人について，ハンドボール投げの飛距離のデータを取った。次の図1は，このクラスで最初に取ったデータのヒストグラムである。

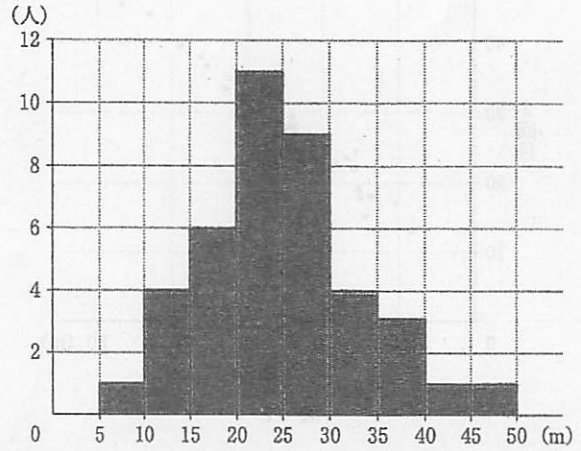


図1 ハンドボール投げ

(1) 次の［ア］に当てはまるものを，下の⓪～⑧のうちから一つ選べ。

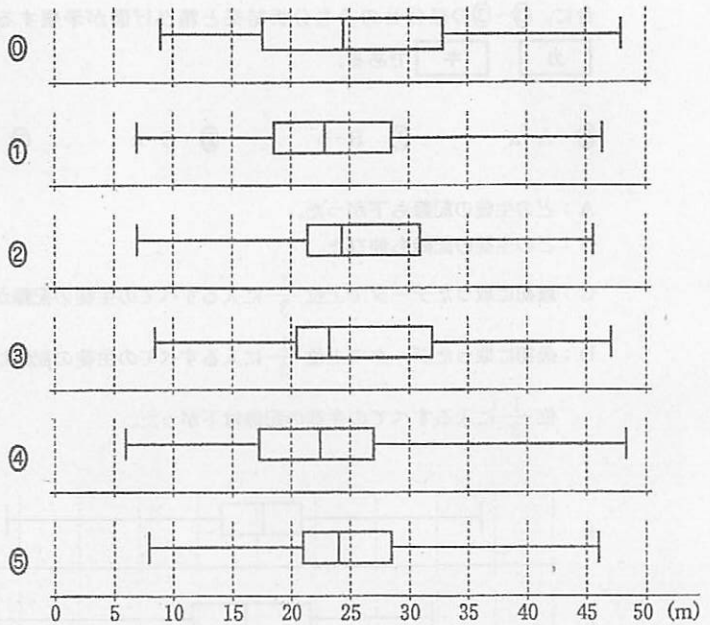
この40人のデータの第3四分位数が含まれる階級は，［ア］である。

⓪ 5 m 以上 10 m 未満
① 10 m 以上 15 m 未満
② 15 m 以上 20 m 未満
③ 20 m 以上 25 m 未満
④ 25 m 以上 30 m 未満
⑤ 30 m 以上 35 m 未満
⑥ 35 m 以上 40 m 未満
⑦ 40 m 以上 45 m 未満
⑧ 45 m 以上 50 m 未満

(数学Ⅰ・数学A第3問は次ページに続く。)

(2) 次の［イ］～［オ］に当てはまるものを，下の⓪～⑤のうちから一つずつ選べ。ただし，［イ］～［オ］の解答の順序は問わない。

このデータを箱ひげ図にまとめたとき，図1のヒストグラムと矛盾するものは，［イ］，［ウ］，［エ］，［オ］である。



(数学Ⅰ・数学A第3問は次ページに続く。)

(3) 次の文章中の［カ］，［キ］に入れるものとして最も適当なものを，下の⓪～③のうちから一つずつ選べ。ただし，［カ］，［キ］の解答の順序は問わない。

後日，このクラスでハンドボール投げの記録を取り直した。次に示したA～Dは，最初に取った記録から今回の記録への変化の分析結果を記述したものである。a～dの各々が今回取り直したデータの箱ひげ図となる場合に，⓪～③の組合せのうち分析結果と箱ひげ図が矛盾するものは，［カ］，［キ］である。

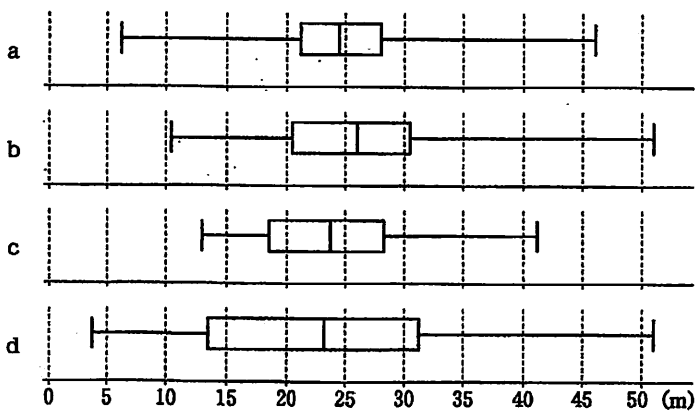
⓪ A-a　① B-b　② C-c　③ D-d

A：どの生徒の記録も下がった。

B：どの生徒の記録も伸びた。

C：最初に取ったデータで上位$\frac{1}{3}$に入るすべての生徒の記録が伸びた。

D：最初に取ったデータで上位$\frac{1}{3}$に入るすべての生徒の記録は伸び，下位$\frac{1}{3}$に入るすべての生徒の記録は下がった。



(数学Ⅰ・数学A第3問は次ページに続く。)

〔2〕 ある高校2年生40人のクラスで一人2回ずつハンドボール投げの飛距離のデータを取ることにした。次の図2は，1回目のデータを横軸に，2回目のデータを縦軸にとった散布図である。なお，一人の生徒が欠席したため，39人のデータとなっている。

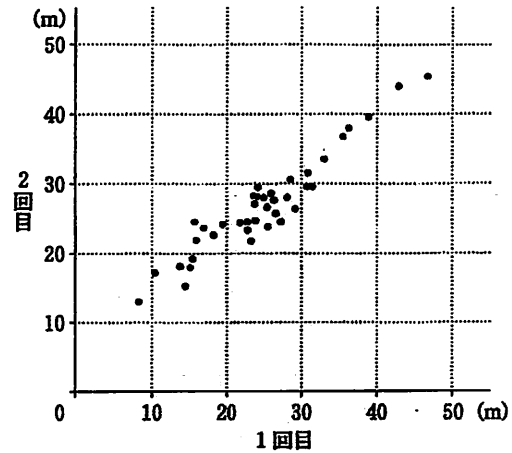


図　2

	平均値	中央値	分　散	標準偏差
1回目のデータ	24.70	24.30	67.40	8.21
2回目のデータ	26.90	26.40	48.72	6.98

1回目のデータと2回目のデータの共分散	54.30

(共分散とは1回目のデータの偏差と2回目のデータの偏差の積の平均である)

次の［ク］に当てはまるものを，下の⓪～⑨のうちから一つ選べ。

1回目のデータと2回目のデータの相関係数に最も近い値は，［ク］である。

⓪ 0.67　① 0.71　② 0.75　③ 0.79　④ 0.83
⑤ 0.87　⑥ 0.91　⑦ 0.95　⑧ 0.99　⑨ 1.03