教師のための統計Tips
Statistical Tips for Teachers

馬　場　　裕

要　旨

本稿では，筆者が教員免状更新講習を現職教員向けに行った際に講義した内容や大学の教員養成課程において講義した内容の中で，統計に関係あることの中から知っておいて欲しいことや間違って認識していることが多いものの一部を抜粋し，加筆・整理したものを解説する．教師のみではなく，大学生や一般人にも役に立つ内容である．

キーワード：四分位数，中学・高校教科書，Excel，相関係数，偏差値

1．はじめに

本稿では教師や大学生（特に教員を目指している者）が知っていて欲しい統計教育に関する話題のいくつかをまとめたものである．筆者は教員免許状更新講習の講師を長く務めてきたが，講習においては教員としてぜひ知っておいて欲しいことや，間違って認識していそうなことを中心に教えてきた．また，大学の教員養成課程においてそれらの内容の一部を講義で取り上げてきた．その中で統計分野で特に興味深いと思ったものをいくつか紹介する．2節では中学・高校の数学教科書における四分位数の定義（文部科学省が学習指導要領に定めたもの）とExcelの四分位数を求める関数の定義が違うため，指導する上で注意したいことを述べた．3節では倍率が高い大学受験等でよく見られる例だが，合格者分布における科目間の相関係数が負になるという不思議な事実がなぜ起こるのかという理由について論じている．4節では偏差値を使用する上でぜひ知っておいて欲しい2つのことについて論じた．一つは偏差値が100点満点のテストのとる値が0から100までの間の値しかとらないわけではないこと，もう一つは各科目の偏差値と合計点の偏差値の間に生ずる興味深い関係である．特に4節（2）で述べた内容を論じた文献は筆者の知る限りでは見たことがないが，この事実は非常に興味深い内容である．

神戸親和大学　教育学部　児童教育学科　教授

2．中学・高校の数学教科書とExcelの四分位数の違い

データを値の大きさの順に並べたとき，4等分する位置にくる値を四分位数という．四分位数は小さい方から第1四分位数，第2四分位数，第3四分位数といい，Q_1，Q_2，Q_3で表す．第2四分位数Q_2は中央値（メジアン）である．日本では学習指導要領により中学・高校の数学教科書において四分位数を次のように定めている．

データを値の小さい方から順に左から並べたとき，左半分のデータを下位のデータ，右半分のデータを上位のデータと呼ぶことにする．ただし，データの大きさが奇数のとき，中央の位置にくる値は，下位のデータにも上位のデータにも含めないものとする．そして，第1四分位数Q_1，第3四分位数Q_3を次で定める．

第1四分位数Q_1は　下位データの中央値

第3四分位数Q_3は　上位データの中央値

具体例として

データA（13個）：10, 20, 20, 20, 40, 40, 40, 50, 60, 70, 70, 70, 80

データB（12個）：10, 10, 10, 30, 30, 40, 40, 60, 60, 80, 80, 80

を見てみる．

データAの中央値（第2四分位数）はデータサイズが奇数のため左から7番目の40である．データAの第1四分位数は7番目の40を除いたデータAの前半半分の10, 20, 20, 20, 40, 40の中央値として，20と20の平均であるから20となる．第3四分位数も同様にして，70と70の平均であるから70となる．

データBの中央値（第2四分位数）はデータサイズが偶数のため左から6番目と7番目の平均の40である．データサイズが偶数のため，下位データの個数は左の6個である．データBの第1四分位数はデータBの前半半分の10, 10, 10, 10, 30, 30, 40の中央値として，10と30の平均であるから20となる．第3四分位数も同様にして70となる．

ところが四分位数は他にもいくつかの定め方があり，教育現場での指導でもよく使われているExcelにおける四分位数を求める関数（2つある）は学習指導要領による定義と違っている．また，RやSPSS等の他の統計ソフトでもExcelとは違う定義をしたものもあり注意を要する．ただし，中央値である第2四分位数Q_2の定義はすべて同じである．

Excelでは四分位数を求める関数としてQUARTILE.EXC（EXCはexclusiveの略）とQUARTILE.INC（INCはinclusiveの略でQURTILEもQUARTILE.INCと同じ）の2つが用意されている．

データC　1，2，3，4，5，6，7，8，9

データD　1，2，3，4，5，6，7，8，9，10

データE　1，2，3，4，5，6，7，8，9，10, 11

データF　1，2，3，4，5，6，7，8，9，10, 11, 12

で与えられる個数の違う 4 つのデータについて教科書およびExcelの 2 つの関数の四分位数の違いを表 1 に示す．

	データC	データD	データE	データF
中・高数学教科書第 1 四分位数	2.5	3	3	3.5
QUARTILE.EXC 第 1 四分位数	2.5	2.75	3	3.25
QUARTILE.INC 第 1 四分位数	3	3.25	3.5	3.75
中・高数学教科書第 3 四分位数	7.5	8	9	9.5
QUARTILE.EXC 第 3 四分位数	7.5	8.25	9	9.75
QUARTILE.INC 第 3 四分位数	7	7.75	8.5	9.25

表 1

データの大きさが大きい場合はどの求め方でも値がほとんど同じになる場合が多いのでそれほど問題にはならないが，授業等では計算のしやすさのためにデータの大きさが小さい場合を扱うことも多いので，指導の際には注意を要する．四分位数の定義は多くあるため，大学入試センター共通テスト（旧センター試験）や統計検定等では四分位数をきっちり求める必要のある問題は出題されていない．

3．受験者分布には正の相関があるが合格者分布には負の相関がある例

大学入試等で 2 科目入試の場合を考える．例えば英語と数学の 2 科目で合否を決める場合とか，大学入試センター共通テスト（旧センター試験）と個別学力試験のそれぞれの総合点を 2 科目の得点と考えてもよい．ほとんどの場合 2 科目の受験者分布には当然ながら正の相関がある．しかし，特に倍率が高い場合には合格者分布には負の相関が見られることがよくある．すなわち，英語がよくできて数学ができない受験生や数学がよくできて英語ができない受験生が合格しやすいというようなことが起こっているように見えるのである．これは決して特殊な事情によるものではないことを解説する．

具体的なシミュレーションとして 2 次元正規分布に従う乱数を発生させて考えてみる．乱数の発生にはExcelの統計マクロ「分析ツール」を使うのが便利である．独立な標準正規乱数，すなわち$N(0, 1)$に従う乱数をX_1，X_2とすると，相関係数ρをもつ 2 次元標準正規分布，すなわち

$$N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$$

に従う2次元標準正規乱数（Y_1，Y_2）は次のように表現できる．

$$Y_1 = X_1$$
$$Y_2 = \rho X_1 + \sqrt{1-\rho^2}X_2$$

さらに，

$$Z_1 = \mu_1 + \sigma_1 Y_1$$
$$Z_2 = \mu_2 + \sigma_2 Y_2$$

とすれば，(Z_1, Z_2) は 2 次元正規分布

$$N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$$

に従い，その確率密度関数は

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right\}$$

である．

2 科目とも平均50点，標準偏差15点で相関係数が0.4の 2 次元正規乱数を200個発生させ（受験人数200人の 2 科目入試と考える），それを散布図にしたのが図 1 である（200個の 2 次元正規乱数による標本相関係数は0.37である）．もし受験倍率が 5 倍で合格者を40人とした場合，2 科目の得点をそれぞれ x，y，上から40番目の総合点を k とすると，$x + y \geqq k$ を満たしたときに合格となる．すなわち傾き -1 の直線で合格ラインを引くことになる．合格者40人の得点の散布図を図 2 に示すと標本相関係数は -0.42 というかなり大きい負の値となっていることがわかる．倍率が高ければ高いほど合格ラインが上の方になるので，相関係数が負でかつその絶対値も大きくなることがあり得ることがわかるであろう．さらに，入学者のみの相関係数は滑り止めで受験している超上位層が抜けるため，さらに絶対値が大きくなることが多い実例がある．ここでは 2 科目の例で示したが，受験科目が 3 科目以上ある場合，2 つずつの科目で相関分析を行えば，同じような結果が得られる場合も多くみられている．

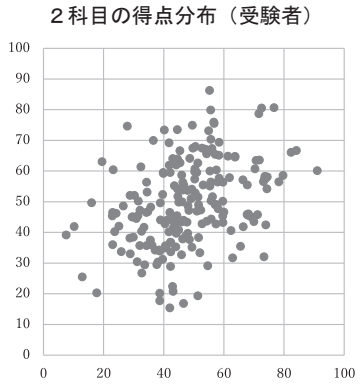

図 1

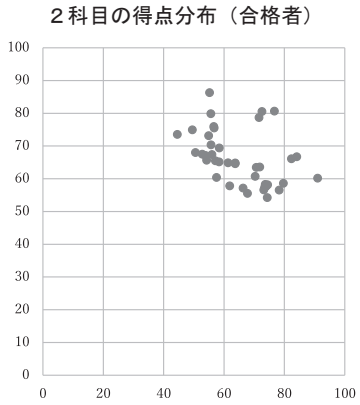

図２

４．偏差値の不思議

この節では偏差値に関わる興味深い２つの話題について紹介する．

（１）偏差値のとる値は０から100の間に収まるか？

偏差値はデータを平均50，標準偏差10のデータに変換（標準化）ものである．具体的にはデータサイズnのデータを$x_1, x_2, \ldots, x_n$とすると，

平均値　$\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

分散　$s^2 = \dfrac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$

標準偏差　$s = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2}$

を使って，x_iの偏差値T_iは次の式で定義される．

$$T_i = \frac{10(x_i - \bar{x})}{s} + 50$$

実際に偏差値を使う際には，異なるデータでの偏差値の比較はデータが正規分布に近いことを前提としている．したがって，データが正規分布に大きく従わない場合は偏差値は必ずしも適切な指標とはいえないことは忘れてはならない．

では偏差値T_iがとる値はどうなるであろうか．偏差値を100点満点のテストの点数を平均を50点とした変形版みたいなものと思って，偏差値のとる値は０から100の間にあると思っている人が少なからず存在する．一般的な100点満点のテストでは100点をとっても偏差値は75から80くらい，０点をとっても偏差値は20から25くらいのことが多い．例えば，平均50点，標準偏差20点のテストでの偏差値は100点で75，０点で25となる．しかし，平均が50点でも標準偏差が５点の場合は，定義からわかるように100点の偏差値は150，０点の偏差値は－50と

なる．また，正規分布からははるかにかけ離れたデータだが，100人が受けたテストで0点が99人で1人だけ100点だった場合は，0点の99人の偏差値は－49，100点の1人の偏差値は149.5となる．このように，偏差値は100点満点のテストの点数の変形版ではないので，必ずしも0から100の間の値をとるわけではない．

（2）各科目の点数の偏差値と合計点の偏差値の関係

例えば，国語，数学，英語の3科目のテストを受験してそれぞれの偏差値がすべて60だったとする．このとき，3科目の合計点の偏差値は60だろうか？この質問を多くの人にしてみたところ，「必ず60になる」とか，「60より小さくなったりちょうど60になったり60より大きくなったりする場合がある」との回答がほとんどであった．実はこの問題の正解は「必ず60より大きくなる」である．これについて解説する．

n人がk科目のテストを受験したとき，番号iの受験生の科目jの得点を

x_{ij}（$i=1, 2, \ldots, n$ ；$j=1, 2, \ldots, k$）とする．

科目jの平均　$\bar{x}_j = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_{ij}$　　（$j = 1, 2, \ldots, k$）

科目jの分散　$s_j^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$　　（$j = 1, 2, \ldots, k$）

科目ℓと科目mの共分散　$s_{\ell m} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} (x_{i\ell} - \bar{x}_\ell)(x_{im} - \bar{x}_m)$　　（$1 \leqq \ell < m \leqq k$）

番号iの受験生の合計点　$y_i = \displaystyle\sum_{j=1}^{k} x_{ij}$　　（$i = 1, 2, \ldots, n$）

とすると，合計点の平均$\bar{y}$と分散s_y^2は次式で表される．

$$\bar{y} = \sum_{j=1}^{k} \bar{x}_j$$

$$s_y^2 = \sum_{j=1}^{k} s_j^2 + 2 \sum_{1 \leqq \ell < m \leqq k} s_{\ell m}$$

簡単な例として，$k = 2$，$s_1^2 = s_2^2 (= s^2)$，科目1と科目2の相関係数が$\rho = \dfrac{s_{12}}{s_1 s_2} = 0.4$のとき，2科目とも偏差値60の場合の合計点の偏差値を計算してみる．

$$\frac{\bar{x}_{i1} - \bar{x}_1}{s} = \frac{\bar{x}_{i2} - \bar{x}_2}{s} = 1$$

$$s_y^2 = s^2 + s^2 + 2s_{12} = s^2 + s^2 + 2 \times 0.4 s^2 = 2.8 s^2$$

より，

$$\frac{y_i - \bar{y}}{s_y} = \frac{s + \bar{x}_1 + s + \bar{x}_2 - \bar{x}_1 - \bar{x}_2}{\sqrt{2.8}s} = \frac{2}{\sqrt{2.8}} \fallingdotseq 1.20$$

だから，番号iの受験生の合計点の偏差値はおよそ

$$50 + 10 \times 1.20 = 62.0$$

となる．s_y^2 の式からわかるように，科目間の相関係数がもっと小さければ合計点の偏差値はもっと上がる．厳密にいえば，科目間の相関係数がすべて１のときは，各科目の偏差値がすべて60ならば合計点の偏差値も60になるが，そんなテストは現実的に存在しないので，60より大きくなると言ってよい．また，各科目の偏差値がすべて50ならば，合計点の偏差値も50となるのは明らかであり，さらに，各科目の偏差値がすべて50未満の場合，例えばすべて45の場合は合計点の偏差値は45未満となる．実際にはすべての科目の偏差値が同じということはなかなかないが，すべての科目の偏差値が50より大き（小さ）ければ，合計点の偏差値は各科目の偏差値の平均よりもほとんどの場合大きい（小さい）．

５．おわりに

本稿では筆者が教員免許状更新講習や大学における統計分野の講義の中で，間違って認識しやすいことやぜひ知っておいて欲しいことのいくつかを紹介した．統計分野だけでも他にも解説したいこともたくさんあり，さらに他の数学の内容においてもあまり教科書や講義等では触れられていないが大切なことが多くある．これらのことについて今後まとめて紹介していきたいと考えている．

参考文献

［１］教師のための統計入門，福島県教育センター，1980.

［２］中学校学習指導要領（平成29年告示）解説　数学編，文部科学省，2017.

［３］高等学校学習指導要領（平成30年告示）解説　数学編理数編，文部科学省，2018.

［４］高等学校数学科用　文部科学省検定済教科書　数Ⅰ／712　数研出版，2021.