全国学力テストを有効活用する「平均ゾーンシステム」の新規開発

―Z 検定と効果量の可視化―

*田　端　健　人・　**丸　山　千佳子・　***本　図　愛　実・
****原　田　信　之・　*****野　坂　実　央

要　旨

全国学力テストを有効活用するために、私たちは「平均ゾーン」という新概念を創作し、各学校・教育委員会の平均正答数と分布曲線が、全国からどの程度ズレているかを可視化する「平均ゾーンシステム」を新規開発した。本稿では、この新システムを、実用的ならびに理論的に解説する。1章で、この開発を導いた問題意識を述べる。一般に全国平均と各学校平均との数値のひらきが広く話題になるが、数値のひらきが「大きい」か「小さい」かの基準は不明のままであった。これが本稿の問題意識である。2章では、本システムの仕組みと出力画面を示し、グラフの見方を解説した。3章では、本システムの理解を促すために、活用例として学校間の比較を示した。4章では異なる活用例として、ある自治体の平成30年度と令和3年度の同一集団の経年変化を可視化した。5章では、この平均ゾーンシステムが前提とする2つの仮説を示した。第1仮説は「47都道府県の平均値に、実質的な差はない」であり、これを本稿では、全都道府県の平均正答数の分布曲線と箱ひげ図により可視化した。第2仮説は「約30名以上の集団の学力分布の、正規分布からの逸脱は、統計的に無視できる」であり、これをF小学校の児童集団29名のデータをK-S検定することで実証した。6章では、本システムの理論背景を述べた。本システムは、Z検定と効果量を直感的に可視化するものである。Z検定は、統計的有意差の有無を基準とし、全国と標本との平均差の「大きさ」を評価する。しかし、z 値や p 値はサンプルサイズに影響され、サンプルサイズが大きければ効果が小さくても有意差ありになる、という問題を指摘した。そこでサンプルサイズに影響されない効果量を合わせて提示する必要がある。ところが、効果量にしても、基準値の難問が浮上し、基準値設定には、主観的判断が避けられない。この難問を解消するために考案したのが、本システムである。7章では、ある小学校の全国学力テスト結果を、7年間追跡した活用例を紹介する。最後の8章で、今後の課題と展望を述べた。私たち「子ども教育データサイエンスDS-EFA」チームの目標は、数量的エビデンスと実践感覚とを架橋し、学校教育の質向上を図り、すべての子どもと社会のウェルビーイングを高めることにある。

Key words：平均、標準偏差、正規分布曲線、p値、活用例、47都道府県の平均差、現職教育

*　宮城教育大学 教職教育総合学域 教育科学部門（教育学）
**　宮城教育大学 高度教職実践専攻運営委員会（学校経営）
***　宮城教育大学 教職教育総合学域 教育科学部門（教育行政学）
****　名古屋市立大学
*****　宮城教育大学

1．問題意識

全国学力・学習状況調査（以下「全国学力テスト」）の結果は、教育委員会や学校やメディアで、適切に有効活用されてきただろうか。実施から15年[1]、全国学力テストの結果は、全国平均正答率を目安とし、その数値との差で議論されてきた。しかし、何ポイントの差が「実質的な差」であるかは、これまで明らかにされてこなかった。そのため、実質的な差がないにもかかわらず、その差に過剰反応する論調が流布し続けている。その象徴が、都道府県別ランキングである。

私たちは、先の研究で、K-S 検定、中心極限定理、PISA の長期トレンドに関する OECD の分析などのエビデンスをもとに、「各都道府県の平均値に、実質的な差はない」ことを検証した[2]。

一方、個別の学校や学級サイズで見ると、全国平均を大きく下回る学校・学級、大きく上回る学校・学級が存在するのも事実である。大きく下回る学校・学級は、授業や学習指導の改善により、児童生徒に学力をいっそう保障する課題がある。大きく上回る学校・学級は、その指導方法の効果が評価され、広く共有されるべきである。

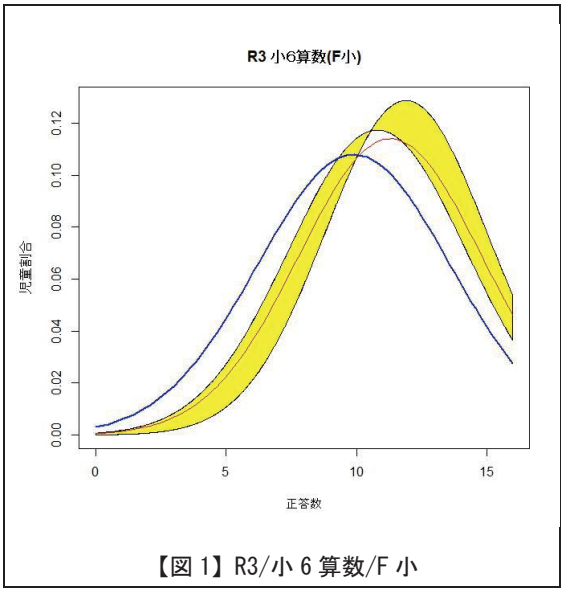
しかし、ここで問題になるのは、「大きく下回る」とか「大きく上回る」とかの「大きく」を、何を基準に判定するかである。全国平均に対してどのくらい差があれば、「大きな差」あるいは「実質的な差」とみなすべきだろうか。

2．「平均ゾーン・システム」の出力

この問題を解決するために、私たちは、全国学力テストの「**平均ゾーン（Average Zone）**」という新概念を考案し、各学校・学級の分布曲線が、この平均ゾーンからどの程度ズレているかを可視化するシステム「**平均ゾーンシステム（Average Zone System）**」を新規開発した。

このシステムは、現在のところ、統計ソフトＲでプログラムされている。全国学力テスト結果から、全国の平均正答数（「AVE.」）と標準偏差（「S.D.」）、トップ都道府県の平均正答数と標準偏差、ボトム都道府県の平均正答数と標準偏差を事前に入力したＲスクリプトに、特定の自治体や学校の平均正答数と標準偏差を入力することで、分布曲線をＲが自動で作図する。

このシステムに、令和３年小６算数Ｆ小学校のデータ（AVE.=9.9、S.D.=3.7）を入力した出力を示すと、図１のようになる。



【図 1】R3/小 6 算数/F 小

横軸は正答数（全16問）、縦軸は児童割合である。

黄色で塗りつぶした領域が、平均ゾーンである。ゾーンの上限カーブはトップ自治体（石川県）、下限カーブはボトム自治体（奈良県）になっている。

赤色のカーブは全国平均の分布曲線である。

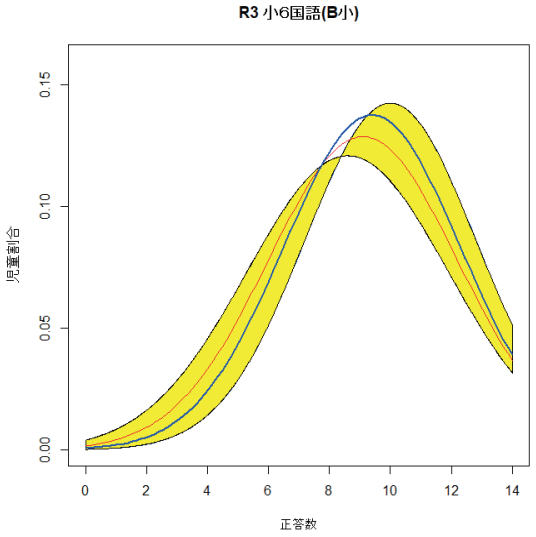
カーブはいずれも、平均正答数と標準偏差から導出される正規分布曲線である。

青色がサンプルのＦ小学校の正規分布曲線である。Ｆ小学校のカーブが平均ゾーンから大きく左にずれ（AVE. が全国より低い）、山の高さが低く傾斜がなだらかであり（S.D. が全国より小さい）、算数に課題があることを、グラフは直感的に示している。

平均ゾーンの雛形（フォーマット）を、私たちは、平成26年から令和３年まで、小学６年国語・算数、中学３年国語・数学で作成し終えている。各学校や自治体の平均正答数と標準偏差を入力すれば、平均ゾーンとの対照で、各年度・各教科の当該学校や自治体の効果や課題が、直感的に可視化できる。

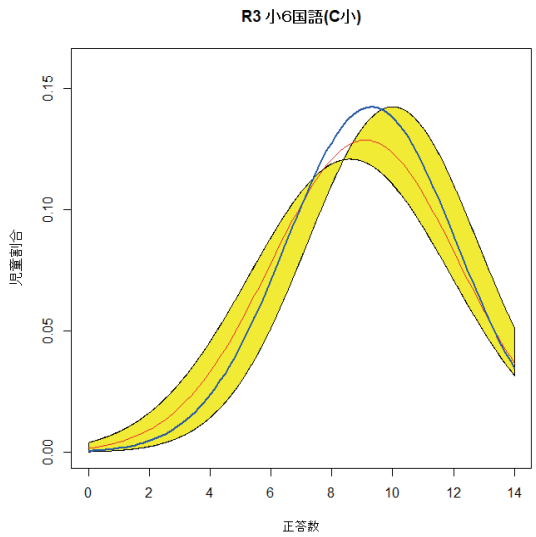
3．平均ゾーンシステムの活用例（１）
―学校間の比較―

本システムの理解を助けるために、活用例を紹介す



全国	AVE. 9.1/14	S.D. 3.1
B 小	9.4/14	2.9

【図 2】R3/小 6 国語/B 小



全国	AVE. 9.1/14	S.D. 3.1
C 小	9.3/14	2.8

【図 3】R3/小 6 国語/C 小

る。まずは学校間の比較である。ただ、競争や序列化を助長する比較は決してしないで欲しい。学校間の比較は、誤った序列化を修正するために利用して欲しい。

例えば、次のＢ小（図２）とＣ小（図３）の比較である。地元では、「B 小は学力が高く、C 小は学力が低い」との評判がある。

しかし、令和３年小６国語で比較すると、B 小と C 小との差はほとんどないばかりか、C 小の方がむしろ良いように見える。

数値としては、平均正答数で14問中、B 小9.4問、C 小9.3問と B 小の方が0.1ポイント高いが、標準偏差で C 小の方が0.1ポイント小さいため、カーブの山が尖り、ゾーンから際立つ形になっている。

「標準偏差が小さい」ということは、「ばらつきが小さい」ことを意味し、多くの場合、低学力児童の底上げができていることを意味する。この例からわかるように、標準偏差の大小も、全国学力テストの結果を評価する基準になる。

従来の議論では平均値だけが話題になるが、平均ゾーンシステムは標準偏差の情報も組み込んでいるため、従来の議論では見逃されていた側面が可視化できる。情報の次元を１つプラスすることになる。

さらに全国平均だけでなく、トップ都道府県とボトム都道府県の情報も入っているため、線ではなく面での評価が可能になる。これで次元が２つプラスになる。

４．平均ゾーンシステムの活用例（２）
―同一集団の経年変化―

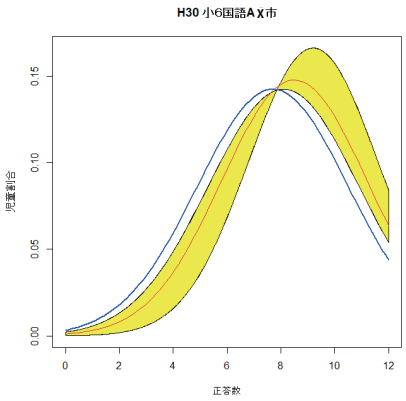
次に同一集団の経年変化を可視化する例を示す。

平成30年度の小学６年生は、令和３年度の中学３年生であるため、自治体規模で見ると、両者はほぼ同一集団とみてよい。全国の同年代の集団も同じである。総人口３万人ほどの X 市の、平成30年小６国語と令和３年中３国語を並べたのが図４・５である。

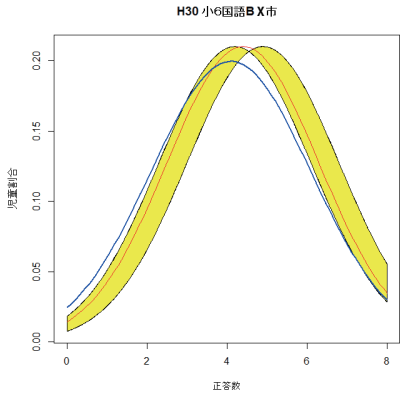
平成30年度は A 問題と B 問題があったため、グラフが２つになり、比較がやや複雑である。しかし、X 市の場合、平成30年小６国語では、A・B 問題とも平均ゾーンから山のすそ野が左にずれ、黄色ゾーンと青曲線との間に余白ができていたのに対し、令和３年では左右のすそ野が平均ゾーンの黄色部分にしっかり入り込んでいる。平成30年小６の生徒集団を、X 市の中学校が学力向上させたエビデンスである。

X 市の集計児童生徒数は、各学年270名前後である。270名もの集団の学力分布を、これほど右にスライドさせるのは、生易しいことではない。

この生徒集団では、算数・数学でも、同様の結果が見られた。X 市の中学校は、数学でも、同様に、この

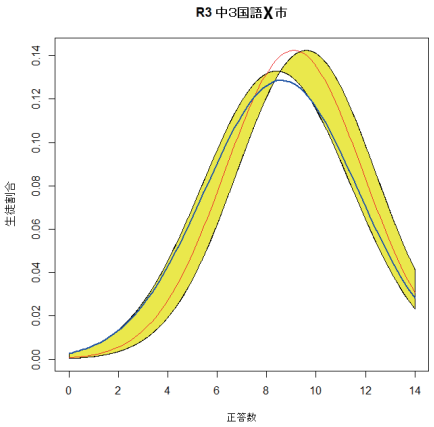


全国	AVE： 8.5/12	S.D: 2.7
X市	7.7/12	2.8



全国	AVE： 4.4/8	S.D: 1.9
X市	4.1/8	2.0

【図 4】H30/小 6 国語 A・B/X 市



全国	AVE： 9.1/14	S.D: 2.8
X市	8.6/14	3.1

【図 5】R3/中 3 国語/X市

生徒集団の学力を向上させている。中学になると国語も数学も内容が高度で難しくなるため、小学校最終学年で平均を下回った児童集団を中学校で押し上げるのは難しい。しかし、X 市の中学校がそれを成し遂げたことを、平均ゾーンシステムは可視化している。

本稿 7 章では、さらに立ち入った活用例を紹介する。

5．平均ゾーンシステムの前提

このシステムは、次の 2 つの仮説を前提としている。

【第 1 仮説】47都道府県の平均値に、実質的な差はない。

【第 2 仮説】約30名以上の集団の学力分布の、正規分布からの逸脱は、統計的に無視できる。

（1）第 1 仮説

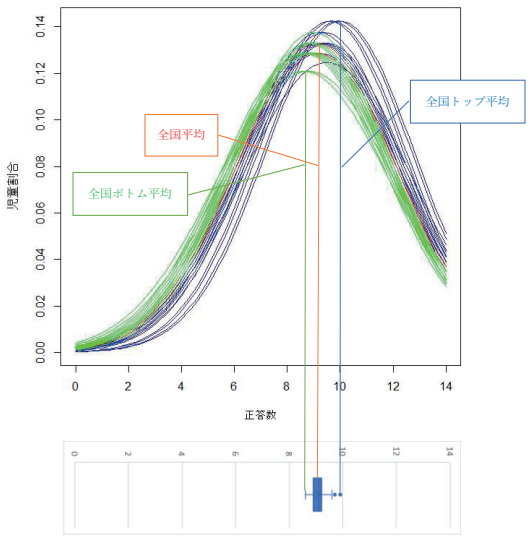
第 1 の仮説について、私たちは先の研究で、K-S 検定、中心極限定理、PISA の長期トレンドに関する OECD の分析などのエビデンスをもとに、「各都道府県の平均値に、実質的な差はない」ことを実証した[3]。

本稿では、47都道府県の平均正答数と標準偏差が、どのような正規分布曲線（カーブ）を描くかを可視化することで、47都道府県の分布と平均値が似通っていることを、直感的に示したい。

まず、令和 3 年小 6 国語の47都道府県の全カーブを作図する。

都道府県は47あるが、平均値と標準偏差のパターンは47もはなく、20前後である。例えば、令和 3 年小 6 国語では、福岡県と宮崎県は、平均正答数9.2、標準偏差3.0とタイであるため、正規分布のカーブも同じになる。

全国平均正答数と同じ平均正答数のカーブを赤、それより高い平均正答数のカーブを青、それより低い平



【図 6】R3/小 6 国語/47 都道府県の正規分布曲線と箱ひげ図

均正答数のカーブを緑で作図した。加えて、その47都道府県の平均正答数の箱ひげ図を横にして併記した（図 6 ）。分布曲線の山の頂点が、その自治体の平均正答数になる。

同様に、令和 3 年の小 6 算数（図 7 ）、中 3 国語（図 8 ）、中 3 数学（図 9 ）の正規分布曲線と箱ひげ図を示す。

これらのグラフから、全国トップ自治体とボトム自治体の平均値差が、いかに小さいかが直感的にわかる。また全都道府県の分布曲線が、いかに近似しているかもわかる。この幅の中で上下を問題にするのは、まったく生産的ではない。それよりも、平均ゾーンとの関係で、各学校や自治体の課題や効果を具体的に把握する方が、よほど生産的である。

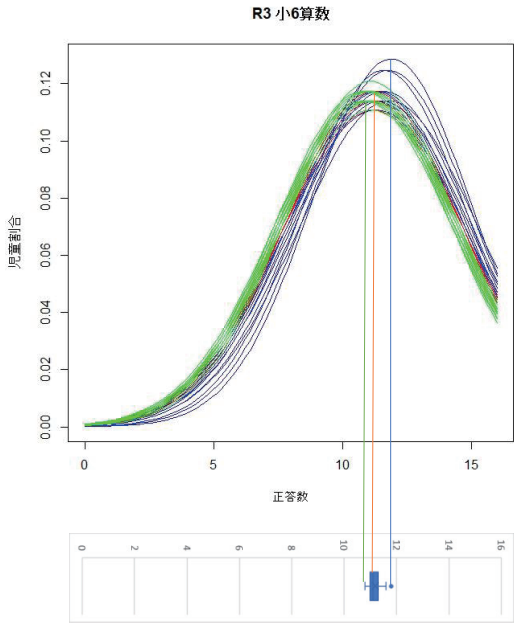
この直感が人びとに共有され、全国学力テスト都道府県別ランキングなどによる過剰な序列化や競争に、終止符が打たれることを期待している。

なお、このカーブの集合体を黄色で塗りつぶし、平均ゾーンとしてもよかったが、そのスクリプトが書けなかったという技術的限界から、頂上付近でねじれのあるバナナの皮を垂らしたような平均ゾーン図となった。頂上付近のねじれ（くびれ）に、特段の意味はない。

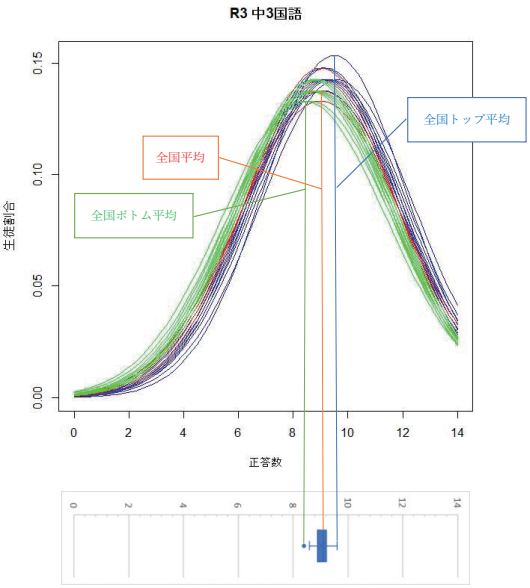
（2）第 2 仮説

第 2 の仮説は、平均ゾーンシステムの理論的背景である Z 検定の前提とも重なる。

以下、第 2 仮説の妥当性を検証する。



【図 7】R3/小 6 算数/47 都道府県の正規分布曲線と箱ひげ図



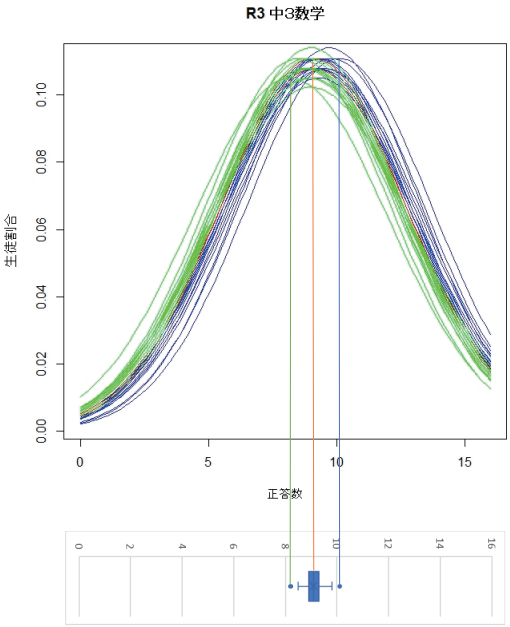
【図 8】R3/中 3 国語/47 都道府県の正規分布曲線と箱ひげ図

検証のため、分布曲線のもととなる、各学校の正答数の分布を、コルモゴロフ・スミルノフ検定（Kolmogorov–Smirnov test；「K-S 検定」）で評価する。

検定の仮説は、以下となる。

帰無仮説（H_0）：標本分布は、正規分布と差がない（正規分布に従う）

対立仮説（H_1）：標本分布は、正規分布と差がある（正規分布に従わない）



【図 9】R3/中３数学/47 都道府県の正規分布曲線と箱ひげ図

有意水準は5% とする。p 値（p-value）< 0.05なら帰無仮説（H_0）が棄却され、標本分布は正規分布から有意にズレていることになる。

p 値の計算には、R の ks.test 関数を用いる。

100名規模の複数の学校の正答数分布を K-S 検定にかけたところ、いずれも p 値 > 0.05で分布の正規性が確認された。

そこで、いっそう小規模の学校を K-S 検定にかけてみる。

以下は、児童数29名の F 小学校小6算数の R スクリプトと計算結果である。

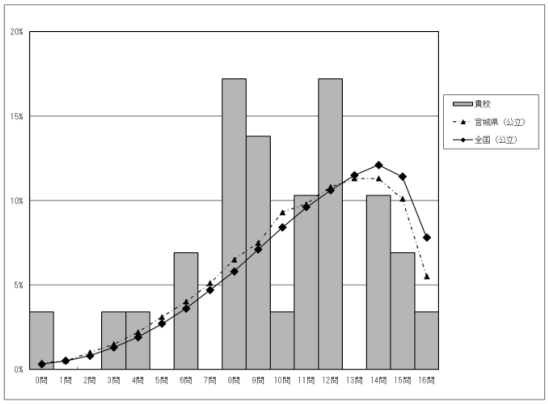
```
>#F 小（29 名）算数_KS 検定
>f_m<-c(3.4, 0.0, 0.0, 3.4, 3.4, 0.0, 6.9, 0.0, 17.2, 13.8, 3.4, 10.3, 17.2, 0.0, 10.3, 6.9, 3.4)
> ks.test(x=f_m, y="pnorm", mean=mean(f_m), sd=sd(f_m))

        One-sample Kolmogorov-Smirnov test

data:  f_m
D = 0.24888, p-value = 0.243
alternative hypothesis: two-sided
```

p = 0.24 > 0.05となり、5% 水準で帰無仮説が保留され、F 小の分布は正規分布に従っているとみなすことができる。

学校にフィードバックされた正答数の分布図は図10である。折れ線グラフが全国と宮城県の分布である。これらはサンプル数が1000をゆうに超える（全国は約100万人）ため、正規分布になる。棒グラフが F 小の実数の割合である。図10を見る限り、分布の正規性は保たれていないようにも見える。しかし検定では、これほどのズレなら統計的に無視でき、正規性は担保されているという結果である。特に３本の棒グラフが正規分布から突出しているが、それは統計的に無視できる、というのが検定のメッセージである。



【図 10】R3/小 6 算数/F 小

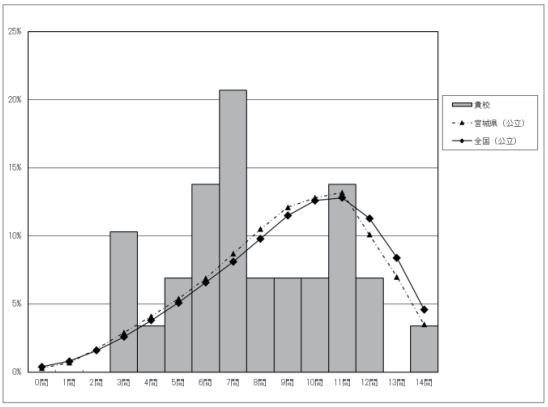
次に、同じく F 小の令和３年国語の分布を K-S 検定してみる。以下が、児童数29名の F 小学校小 6 国語の R スクリプトと計算結果である。

```
>#F 小（29 名）国語_KS 検定
> f_j<-c(0.0, 0.0, 0.0, 10.3, 3.4, 6.9, 13.8, 20.7, 6.9, 6.9, 6.9, 13.8, 6.9, 0.0, 3.4)
> ks.test(x=f_j, y="pnorm", mean=mean(f_j), sd=sd(f_j))

        One-sample Kolmogorov-Smirnov test

data:  f_j
D = 0.21749, p-value = 0.477
alternative hypothesis: two-sided
```

p = 0.48 > 0.05となり、5 % 水準で帰無仮説が保留され、F 小国語の分布の正規分布からの逸脱は、統計的に無視できるとの結果である。こちらも、文部科学省から F 小に返却された結果のグラフを図11に示す。



【図 11】R3/小 6 国語/F 小

これも、抜きん出ている 3 本の棒グラフの正規分布曲線からの逸脱は、統計的には無視してよい、というK-S 検定のメッセージである。

以上から、児童生徒数30名程度以上の集団の正答数分布は、統計的に有意なほどには正規分布から逸脱しないと結論づける。この現象は「正規分布の頑強性」として、統計の専門家に経験的に知られている。

ちなみに、児童数14名の O 小学校、令和 3 年の国語を K-S 検定にかけると、$p = 0.06$となった。5 % 水準で有意ではないが、さすがにこれを正規分布と仮定するには無理がある。

これらの結果から、平均ゾーンシステムは、およそ30名以上の集団の評価には有効であり、それ以下の人数規模の集団の評価には不適切であると結論づける。

もしも30名程度の集団の正規性で、K-S 検定が有意な結果になれば、その集団には、なにか例外的な事情が働いていると考えてよい。

6．平均ゾーンシステムの理論背景

本システムは、Z 検定と効果量の測定を理論背景としている。結論からいえば、本システムは、Z 検定と効果量の直感的可視化である。

（1）Z 検定

Z 検定とは、正規分布の母平均に関する仮説検定で、「母分散σ^2が既知の場合」に用いられる。全国学力テストは悉皆調査である。そのため母分散σ^2が既知となり、Z 検定が使えることになる。Z 検定により、全国学力テストの全国平均値と、標本（サンプル）の学校平均値との差を、統計的に評価できる。

Z 検定には、もう一つ重要な前提がある。「標本データは正規分布 N（μ，σ^2）に従う」という仮定である。この仮定は、先の第 2 仮説の検証において、K-S 検定で担保された。児童生徒数30名程度以上なら、正規分布からのズレは統計的に無視できると仮定でき、平均ゾーンシステムを利用できる。これ以下の人数では、このシステムの評価は有効ではない。

話を戻すと、Z 検定は、標本の学校が全国平均から「統計的に有意にズレているか否か」を評価する。統計的有意差の有無が、全国平均との差の「大きさ」の有無の基準となる。

検定仮説は、次のようになる。μ は標本平均、μ_0 は母平均である。

帰無仮説 H_0：$\mu = \mu_0$（標本平均は全国平均と同等である；全国平均と差は無い）

対立仮説 H_1：$\mu > \mu_0$（標本平均は全国平均と有意に差がある；全国平均との差は大きい）

評価基準は、統計の慣例から 5 % 水準とする。$p <$ 0.05なら、標本平均は全国平均から有意に差がある、つまりその差は「大きい」と評価する。

検定統計量 z の計算式は、次の通りである。σ は母標準偏差（母分散σ^2の平方根）、n は標本数である。

$$z = \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

平成31年小 6 算数で、児童数138名の A 小学校と、児童数188名の B 小学校を Z 検定にかけてみる。事前の K-S 検定で、A 小 B 小の分布の正規性を確認している。

まず全国（国公私立）平均$\mu_0 = 9.3$、母標準偏差$\sigma = 3.1$、A 小平均 $\mu = 10.2$、標本数$n = 138$を代入して計算する。右端の片側検定で、有意水準 5 % の場合、標準正規分布表から$z > 1.64$ならH_0を棄却する。

```
> z <- (10.2 - 9.3)/(3.1/sqrt(138))
> z
[1] 3.410518
```

A 小の場合、$z = 3.41 > 1.64$となり、全国平均より有意に高い、という結果である。

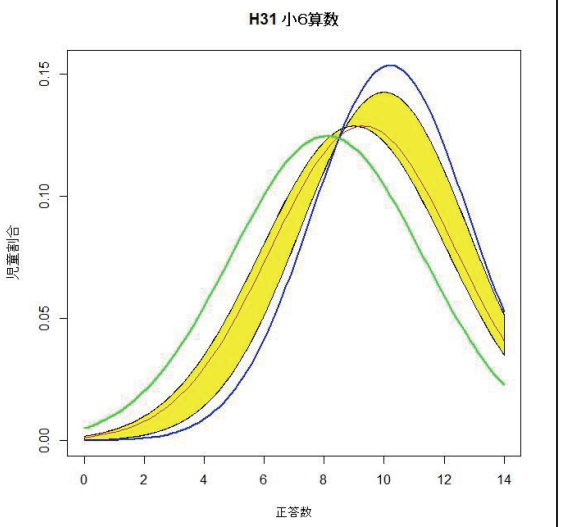
次に B 小平均$\mu = 8.1$、標本数$n = 188$を代入して計算する。左端の片側検定で、有意水準5% の場合、標準正規分布表から$z < -1.64$ならH_0を棄却する。

Rスクリプトと結果は、次の通りである。

```
> z <- (8.1 - 9.3)/(3.1/sqrt(188))
> z
[1] -5.307604
```

B小のZ値 = -5.31 < -1.64となり、H_0は棄却され、B小は全国平均より有意に低い、という結果である。

2つの学校のカーブを平均ゾーンシステムで描くと



【図12】H31/小6算数/A小とB小

図12のようになる。

右側の高い青カーブがA小、左側の低い緑カーブがB小である。全国平均カーブ（赤）と両カーブ（緑・青）とがこれほどズレる場合は、統計的に有意な（大きな）差との評価になる。

（2）効果量

*p*値（p-value）については、誤用と誤解が多く、アメリカ統計協会（American Statistical Association, ASA）は2016年に声明を発表している。声明では次のように述べられている。

「P値はデータと特定の統計モデル（訳註：仮説も統計モデルの要素のひとつ）が矛盾する程度をしめす指標のひとつである。」[4]

例えば、*p*値が0.05より小さければ、帰無仮説H_0（統計モデル）の矛盾が大きいことになり、このモデルを保留できず、対立仮説H_1を採用したほうがよい、という指標の一つとなる。しかし、「科学的な結論や、ビジネス、政策における決定は、P値がある値（訳註：有意水準）を越えたかどうかにのみ基づくべきではない。」[5]とASAは警鐘を鳴らす。

「P値や統計的有意性は、効果の大きさや結果の重要性を意味しない。」[6]

「P値は、それだけでは統計モデルや仮説に関するエビデンスの、よい指標とはならない。」[7]

そこでASAは、「P値以外のアプローチ」を推奨し、「信頼区間、信用区間、予測区間などの、検定よりも推定を強調した方法、ベイズ流の方法、尤度比やベイズファクターなどのことなったエビデンスの指標、そして決定理論やFalse Discovery Rateといったアプローチ」をあげている[8]。

特に、*p*値の計算には、標本のサンプルサイズが影響するため、サンプルサイズの影響を受けない効果量と合わせて報告することが勧められる[9]。

例えば、2グループの平均差のt検定をサンプルサイズを変えてシミュレーションすると、$n = 50$では$p > 0.05$にもかかわらず、$n = 100$では$p < 0.05$になるという結果が示されている[10]。

先のZ検定も、サンプルサイズの影響を受ける。

Z検定の計算式は、

$$z = \frac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

であるが、この式は次のように変形できる。

$$z = \frac{\mu - \mu_0}{\sigma} * \sqrt{n}$$

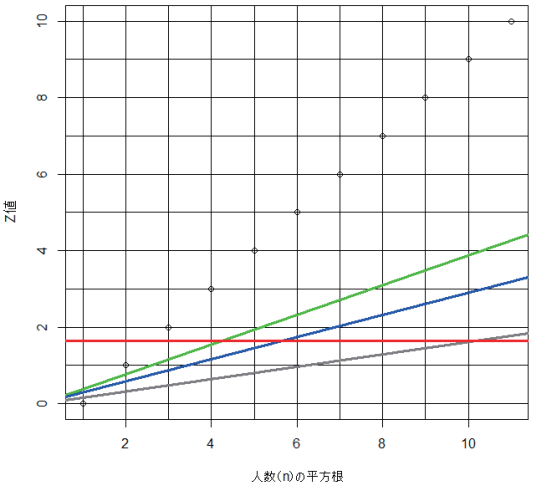
平均差を母標準偏差で割る式（$\frac{\mu - \mu_0}{\sigma}$）は、「効果量」の計算式である。平均差が、標準偏差にしてどのくらい分あるかを、効果として評価している。*n*は人数であるため、人数が多いほど、$\sqrt{n}$が大きくなり、*z*値は大きくなる。そのためサンプルサイズの大きさだけで、5％水準を下回り、有意差ありとなる。つまり、

z値（検定統計量）＝効果の大きさ＊サンプルサイズ

と整理できる。

数式をここまで変形すれば、小さな効果でも、サンプルサイズが多くなれば、Z値が大きくなる、つまり5％水準を下回ることがよくわかる。

平成31年6年算数で、*z*値 = 効果量＊サンプルサイズ（*n*）の一次関数を数直線で表現すると、図13にな



【図 13】*z* 値とサンプルサイズの関係

る。*y* 軸が *z* 値、*x* 軸が *n*（人数）の平方根である。

青線が先の A 小、緑線が先の B 小、灰色線は全国平均との差を0.5と設定したシミュレーションである。赤線が 5 % 水準（1.64）で、この線を越えると有意差ありとなる。緑線は、本来なら、マイナスの傾きの直線となるが、絶対値として扱い、プラスの傾きで表した。

A 小（青色）の全国平均差は0.9、B 小（緑色）の全国平均差は -1.2、シミュレーション（灰色）の全国平均差は0.5である。この平均差を母標準偏差（σ）で割った値が、直線の傾きになる。

A 小は約36人以上のサイズで有意となり、B 小は約16人以上で有意、シミュレーションは約100名以上で有意となる。逆に、A 小の平均差は、A 小の規模が36人未満なら有意差なし、B 小は16人未満の規模なら有意差なし、シミュレーションは100名未満なら有意差なしとなる。

このようにサンプルサイズで有意差の有無が左右されるため、ASA 声明の通り、*p* 値が有意水準を超えたかだけをエビデンスとするべきではない。サンプルサイズの影響を受けない効果量を、合わせてみる必要がある。

効果量は、図13の直線の傾きに対応する。傾きが大きいほど、効果量が大きい。効果量は「*d*」の記号で表される。

A 小の効果量 *d* は、(10.2 - 9.3)/3.1 = 0.29、B 小の効果量 *d* は、(8.1 - 9.3)/3.1 = -0.39となる。シミュレーションの効果量 *d* は、0.5/3.1 = 0.16である。

しかし、ここで難問が発生する。*d* 値の「0.29」とか「-0.39」とか「0.16」という値は、大きいのか、中くらいなのか、小さいのか、どう評価すればよいかという問題である。これは「基準値（目安）」の問題である。

（3）効果量の基準値の問題

結論からいえば、基準値の問題を止揚したのが、新規開発した平均ゾーンシステムである。基準値をどこに設定するかの判断を、データの活用主体にゆだね、ズレの大小を、学校や教育委員会が、直感的・主観的・主体的に判断するよう、発想を転換した。

基準値の問題の難しさを示すために、効果量の基準値に関する議論を紹介しておく。

よく用いられる Cohen（1988）では、*d* = 0.2が効果量小、*d* = 0.5が効果量中、*d* = 0.8が効果量大とされる[11]。

加えて、研究分野ごとに基準を設けるのがよいとの考えもあり、外国語教育研究の分野では、Plonsky and Oswald(2014) が、対応なしのデータで、*d* = 0.40（効果量小）、*d* = 0.70（効果量中）、*d* = 1.00（効果量大）、また対応ありのデータで、*d* = 0.60（効果量小）、*d* = 1.00（効果量中）、*d* = 1.40（効果量大）を提唱している[12]。

（4）基準値に関するハッティの見解

さらに、学習の効果量を可視化し、世界的に注目されているハッティ（Hattie, J.）は、メタ分析研究の結果から、効果量の平均値 *d* = 0.40を基準値と定めている。『学習に何が最も効果的か』刊行時点で、収集された913のメタ分析研究の効果量の分布は正規曲線を描いているが、効果量に着目すると、全要因のうちの95パーセントがゼロ以上のプラス効果を示した。そこで「基準値」が必要となり、そのためにハッティは、学習者、家庭、学校、教師、カリキュラム、授業という 5 つのカテゴリーの効果量の平均値を求めた。その結果が表 1 である。

この結果から、平均的な効果量 *d* = 0.4を、ハッティは基準値とした。ハッティによると、「*d* = 0.4は 1 年での平均的成長として私たちが期待しうるもの」[14]であり、0.4よりも大きな効果量は効果大である。

しかし、全国学力テストの効果量の場合、A 小の効果量を先に計算した0.29とするなら、ハッティの平均0.4に届かない。A 小は全国平均を上回っているの

【表１】５領域の効果量（d）平均値[13]

全領域	メタ分析総数	研究総数	学習者総数	効果指標総数	d	SE
学習者	152	11,909	9,397,859	40,197	0.39	0.044
家庭	40	2,347	12,066,705	6,031	0.31	0.053
学校	115	4,688	4,613,129	15,536	0.23	0.072
教師	41	2,452	2,407,527	6,014	0.47	0.054
カリキュラム	153	10,129	7,555,134	32,367	0.45	0.075
授業	412	28,642	52,611,720	59,909	0.43	0.070
平均	**913**	**60,167**	**88,652,074**	**160,054**	**0.40**	**0.061**

に、ハッティのいう平均的成長0.4を下回るのは理屈に合わない。

では、全国平均の効果を0.4とし、A小の効果量0.29を上乗せし、0.69をA小の効果とみなすべきだろうか。しかし、同じ計算をB小に適用するなら、B小の効果量は0.01になってしまう。確かにB小のテスト結果は芳しくないが、効果量0.01という評価は受け入れ難い。

このように、基準値はどのように設定したとしても、異論や反論や齟齬を免れない。そもそも基準値は、「主観的な推定と仮定の要素がきわめて大きく関与している」[15]。基準値は、本質的にこの宿命を逃れられない。

そこで発想を転換し、ズレの目安や基準の判断を、学校や教育委員会などデータ活用主体にゆだねたのが、平均ゾーンシステムである。このシステムは、学校や自治体の効果や課題を判断するための客観的な分析結果をグラフで提示する。ズレを大きいとみるか、小さいとみるかは、各学校、各教育委員会で議論し、合意形成してほしい。学力の目標を高く設定する学校は、平均ゾーンに入っているだけでは満足しないだろうし、それで良しとする学校・自治体もあるだろう。

７．平均ゾーンシステムの活用例（３）
―A小学校７年の経年変化―

最後に、システムの活用例をもう一つ、やや詳しく紹介する。

参考資料のグラフ一覧は、Y自治体A小学校の平成26年度から令和３年までを経年で追っている。グラフが示す向上は、まさにA小B校長の学校経営と軌を一にしている。B校長は平成26年度にA小に赴任、平成29年度までA小校長を務めた。B校長の取り組みの成果は、赴任１年後の平成27年４月のテスト結果に早くも表れている。

A小が属するY自治体のD教育長（平成24年11月～平成30年度）は、前教育長の「心を育てる」を踏まえつつも、義務教育としての学力向上重視も含むものとして拡充を図り、算数指導で実績のあるB校長をA小に配置し、学力向上の実現を託した。B校長が着任後まず重視したのは、言語活動の充実であった。これは、着任前から校内研究が国語の指導充実をテーマに進んでいたこともあり、その成果が、多様なステイクホルダーからみても明らかになることをめざした。

具体的には、①全校統一的に、家庭学習として音読活動を課した。保護者にも、様々な機会を捉え、親の前で子どもたちが宿題として音読をすること、音読の様子を必ず褒めてほしい、と依頼した。②暗唱大会を実施した。D教育長の支援もあり、自治体の国語研究会を中心に、自治体独自の暗唱読本が作成され、保護者他、地域の関係者にも配布された。暗唱読本を基に、1年から6年までの子どもたちが全校の前で暗唱する様子は、学校内外から賞賛のまととなった。③自主的な読書活動が広がるよう、図書室来室を活発化させた。本の紹介や掲示に工夫を凝らし、お化け屋敷コーナーを作ってミステリーものを置くといった趣向が凝らされることもあった。こうした広範囲に渡る活動が、国語のテスト結果上昇に結びついた。

言語活用能力向上の一方で、B校長の本領発揮ともいえる算数の授業改善が推進されていった。民間テスト、全国学力テストとともに、課題となった点と授業での教え方が分析され、B校長自ら授業を行ない改善

案を提案することもあった。こうした活動は、校内共同研究にも発展し、公開研究会として県内からの参観者の前で授業提案がなされるようにもなっていった。公開研究会では、研究成果の有効性について、教育大学の算数教育研究者も後押しした。

授業改善だけでなく、学習の定着という点では、会議の精選により水曜放課後の一時間が捻出され、子どもたちの個別学習時間として、基本問題の練習が全校で実施された。

D教育長は、県の算数チャレンジ・数学オリンピックへの出場を促し、地域全体で学力向上を後押しする雰囲気を作ろうともした。A小の子どもたちもチームで出場し、上位入賞などの活躍が地元メディアで報じられるようになった。

A小の学力向上において見逃してはならないのが、B校長による、教員への高い期待である。教員の高い動機づけによって、学力向上が実現した。算数だけでなく、全教科について、B校長は、「授業で先生方を褒める」という姿勢に徹し、毎日、同じ時間を先生方の授業を見る時間と決めて各教室を見て回り、教員の授業の工夫について子どもの姿を通して称賛し、あるいは課題の改善について助言した。つまり、授業を行なう教員にとってB校長は、目の前の子どもの学習の改善について共に考え、励ましてくれる存在であった。改善につながる具体的な助言が得られるだけでなく、日々の子どもの姿のわずかな変容に対しても、教員の指導力向上として認められるという仕組みが採られていた。

B校長の指導力向上のためのリーダーシップはまた、D教育長からの高い期待に確実に応えるという、安定的な好循環の中でさらに高い効果を生み出したとも言えるのである。加えて、平成31年のA小の順調な成果についても、D教育長の影響力の可能性が考えられる。

8．今後の課題と展望

私たちの今後の課題は、このシステムを、Webアプリ化し、オンライン上で、誰もが無料で利用できるようにすることである。自分の学校の平均正答数と標準偏差を入力すれば、グラフが描画され、自校のテスト結果を評価できるWebシステムである。私たちDS-EFAチームは、すでにその試作版を完成させている。

これにより、統計やプログラミングを専門としない教員でもデータ分析が可能になる。本稿は、全国学力テストの結果を現職教員が、いっそう効果的に分析・活用することに資することをめざしている。この平均ゾーンシステムがWebアプリ化されるならば、この目標へと大きく近づくであろう。

学校現場や教育委員会が、データ分析の主体となり、実践感覚をデータで裏付けながら、ときにデータに裏切られながら実践感覚を磨き、日本の学校教育の質をいっそう高めてほしい。特に学力向上を実現した学校の効果的な取組みを特定し、その効果の恩恵をすべての子どもに浴させてほしい。学力向上は、非認知能力の豊富化と絡み合っている。非認知能力を土台とした学力向上は、子どもの幸福度を高めることになる。これが、本プロジェクトの大もとにある「子ども教育データサイエンスDS-EFA（ディーエス・イーファ）」のコンセプトである。

野球やサッカーや柔道など、スポーツの分野では、トレーニングでも本番でも、多種多様なデータサイエンスが活用されている。対照的に、子どもの各種能力の成長に責任をもつ学校教育では、データサイエンスはまだほとんど実用されていない。DS-EFAは、データサイエンスにより、こうした現場に変革をもたらし、学校と社会のウェルビーイングの向上をめざしている。

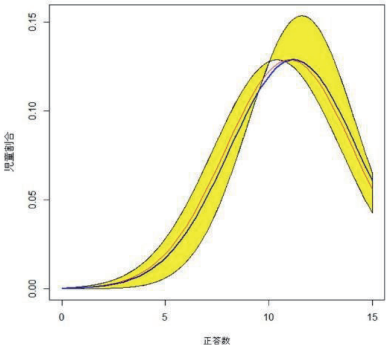
【付記1】全国学力テストで全国平均と各学校平均との差をZ検定で評価する、という画期的着想は、東北大学大学院教育学研究科 柴山直 教授からご教示いただいた。この場をお借りして、心より感謝申し上げる。

【付記2】現職教育充実の観点から、学力データを貸与くださり、自治体や学校や生徒の情報を保護した上で、公開をご快諾くださった自治体や学校の皆様に、心より感謝申し上げる。

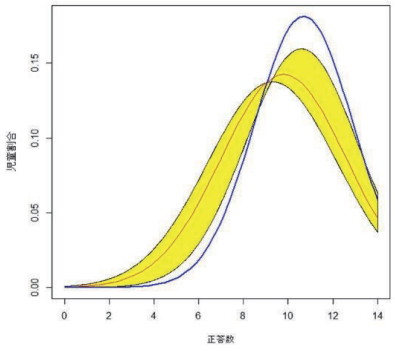
【付記3】本研究は、科学研究費助成事業、基盤研究B「グローバル世界を視野とする学力・非認知能力の効果的学校モデル」(研究代表：田端健人)(2020～2022年度、課題番号：20H01667）の研究成果の一部である。

【参考資料】A 自治体 A 小学校の平成 26 年度～令和 3 年度の経年変化

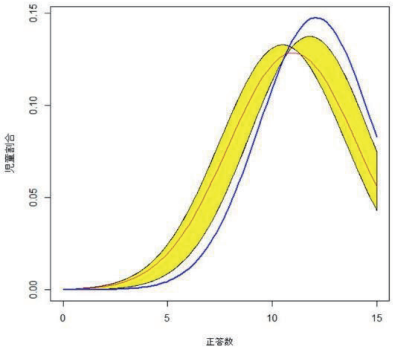
H26/小 6/国語 A



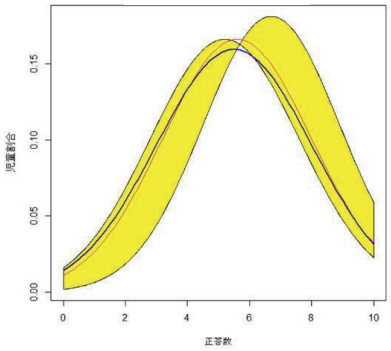
H27/小 6/国語 A



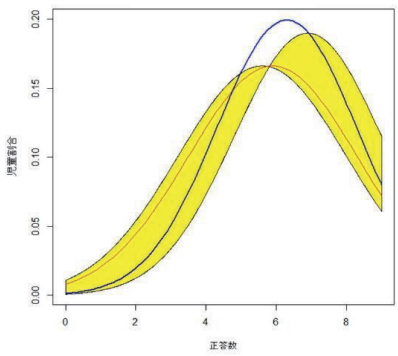
H28/小 6/国語 A



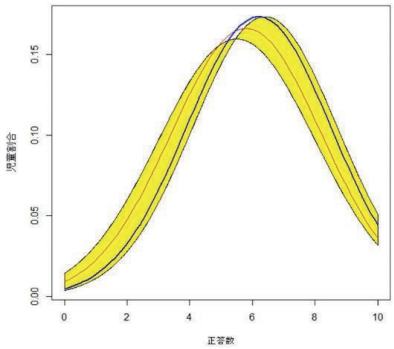
H26/小 6/国語 B



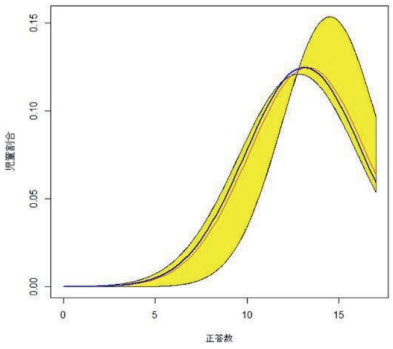
H27/小 6/国語 B



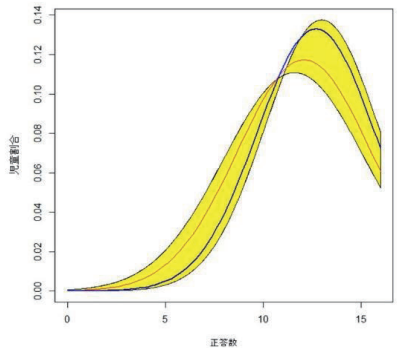
H28/小 6/国語 B



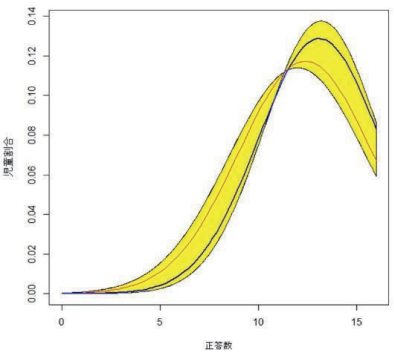
H26/小 6/算数 A



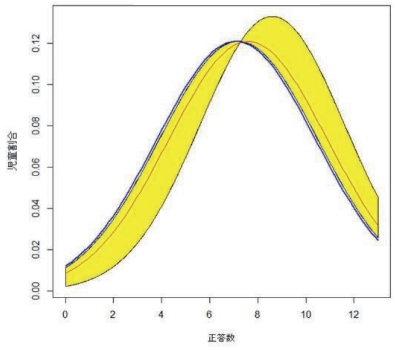
H27/小 6/算数 A



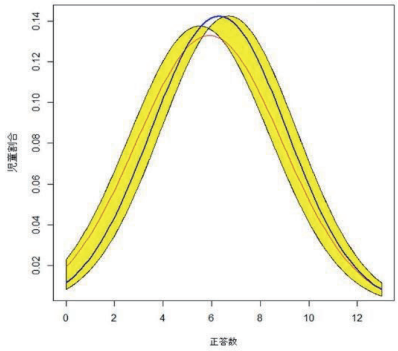
H28/小 6/算数 A



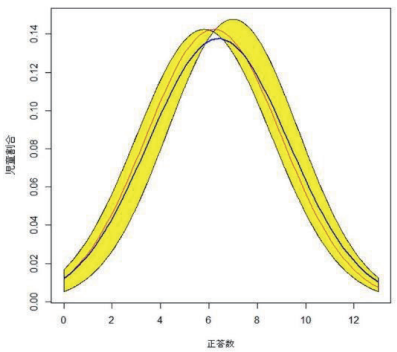
H26/小 6/算数 B



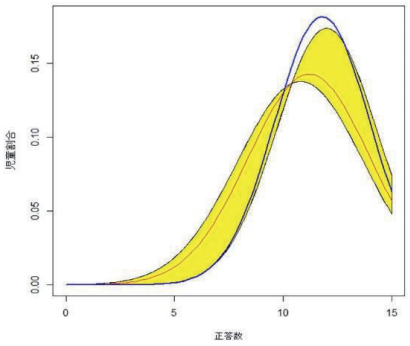
H27/小 6/算数 B



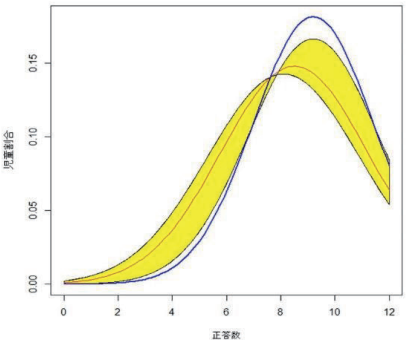
H28/小 6/算数 B



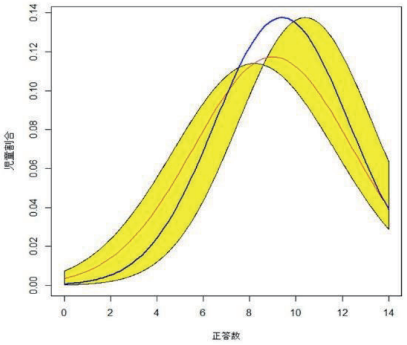
H29/小 6/国語 A



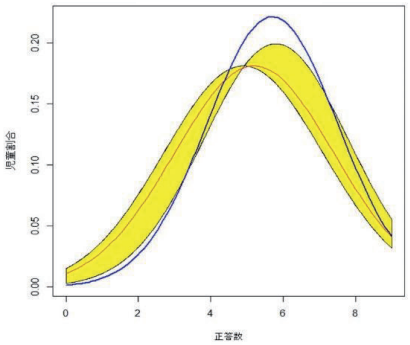
H30/小 6/国語 A



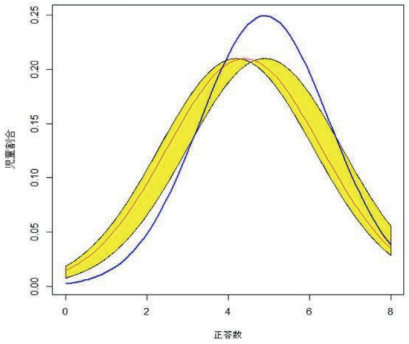
H31/小 6/国語



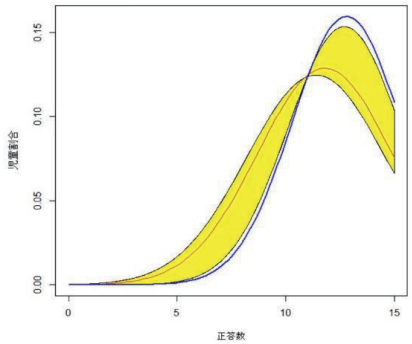
H29/小 6/国語 B



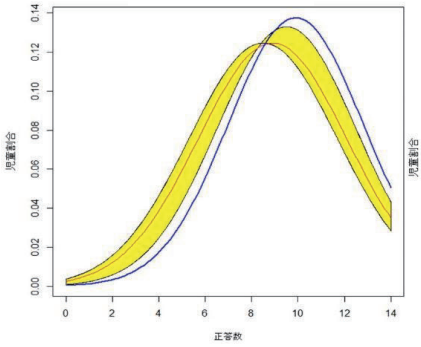
H30/小 6/国語 B



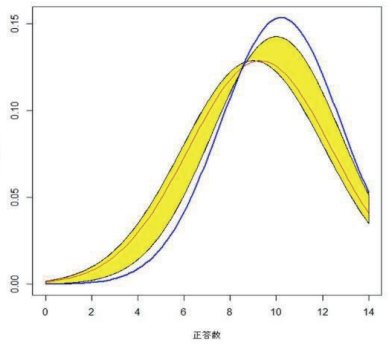
H29/小 6/算数 A



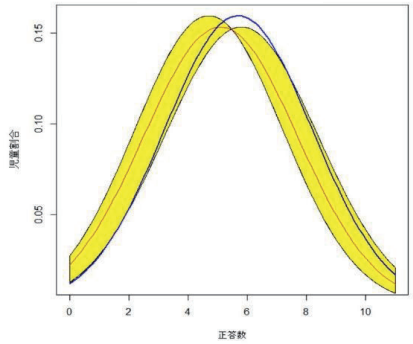
H30/小 6/算数 A



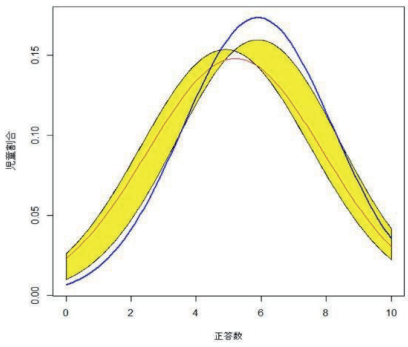
H31/小 6/算数



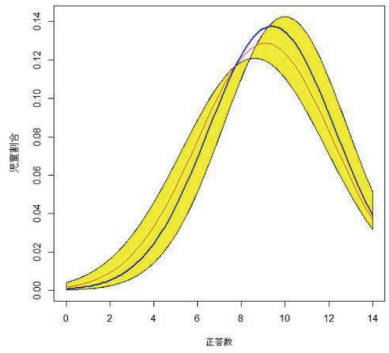
H29/小 6/算数 B



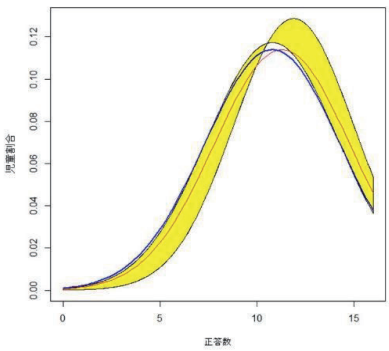
H30/小 6/算数 B



R3/小 6/国語



R3/小 6/算数



注

1　現行の全国学力テストは、2007（平成19）年度にスタートした。

2　田端健人（2021)「全国学力・学習状況調査の平均正答率をどう受けとめるべきか？―『生きられた数値』による＜子ども教育データサイエンス＞の構想―」『学ぶと教えるの現象学研究』19, pp.1 -20.

3　注 2 参照 .

4　佐藤俊哉訳（2017)「統計的有意差と P 値に関する ASA 声明」(PDF 資料)，p.1. https://www.biometrics.gr.jp/news/all/ASA.pdf

5　佐藤（2017), p.2.

6　佐藤（2017), p.2.

7　佐藤（2017), p.3.

8　佐藤（2017), p.3.

9　Cf., 水本篤・竹内理（2010)「効果量と検定力分析入門―統計的検定を正しく使うために―」『より良い外国語教育研究のための方法』, p.48. 小林雄一郎・濱田彰・水本篤（2020)『R による教育データ分析入門』オーム社 , p.90.

10　Cf., 水本ほか（2010), p.48.

11　Cf., 小林雄一郎・濱田彰・水本篤（2020)『R による教育データ分析入門』オーム社 , p. 91.

12　Cf., 小林ほか（2020), p.93.

13　ハッティ , J.（2017)『学習に何が最も効果的か』(原田信之訳者代表）あいり出版 , p.15.

14　ハッティ (2017), p.18.

15　村上道夫・永井孝志・小野恭子・岸本充生（2014)『基準値のからくり―安全はこうして数字になった―』講談社 , p.16.

（令和 3 年 9 月30日受理）

An Invention of “Average Zone System” to Utilize the National Academic Achievement Test

The Visualization of Z-Test and Effect Size

TABATA Taketo, MARUYAMA Chikako, HONZU Manami,
HARADA Nobuyuki and NOZAKA Mio

Abstract

In order to make effective use of the national academic achievement test, we created a new concept called "average zone" and invented the "average zone system". This system visualizes how much the average number of correct answers and the distribution curve of each school deviate from the whole country. In this paper, we will explain this new system practically and theoretically. Chapter 1 describes the awareness of the issues that led to this development. Generally, the average gap between the whole country and each school is widely talked about, but the criteria for whether the gap is "large" remained unclear. This is the problem awareness of this paper. Chapter 2 showed the mechanism of this system and the output screen, and explained how to read the graph. In Chapter 3, in order to promote understanding of this system, a comparison between schools is shown as an example of utilization. Chapter 4 shows another usage example. We visualized the secular change of the same group in a certain municipality. In Chapter 5, we have presented two hypotheses that this average zone system presupposes. Hypothesis 1 is "there is no substantial difference in the average values of 47 prefectures". In this paper, this is visualized by the distribution curve of all prefectures and the boxplot of the average number of correct answers. Hypothesis 2 is that "the academic ability distribution of a group of about 30 or more deviates from the normal distribution, but the deviation is statistically negligible." This was demonstrated by K-S testing data from a group of 29 F elementary school children. Chapter 6 described the theoretical background of this system. This system intuitively visualizes the Z-test and effect size. The Z-test can evaluate the "magnitude" of the mean difference between the whole country and the sample based on the presence or absence of statistically significant difference. However, the z-value and p-value are affected by the sample size, and if the sample size is large, there will be a significant difference even if the effect is small. Therefore, it is necessary to present the effect size that is not affected by the sample size. However, even with effect sizes, the difficult problem of reference values has emerged, and subjective judgment is inevitable when setting reference values. This system was devised to solve this difficult problem. Chapter 7 introduces an example of using the results of a national academic achievement test at an elementary school for 7 years. In Chapter 8, we described future issues and prospects. The goal of our "Data Science of Education for All (DS-EFA)" team is to bridge quantitative evidence and a sense of practice, improve the quality of school education, and enhance the well-being of all children and society.

Key words： Mean, standard deviation, normal distribution curve, p-value, application example, mean difference of 47 prefectures, teacher training