# 文系学生の統計分野の能力分析及び誤答の考察
## ―― 古典的テスト理論及び項目反応理論による結果を踏まえて ――

佐々木　　淳

目　次



## 1. はじめに

内閣府は、AI 戦略 2022（内閣府，2022）において「文理を問わず、全ての大学・高専生（約 50 万人卒／年）が課程にて初級レベルの数理・データサイエンス・AI を習得」する具体的目標を掲げている。掲げられた初級レベルの教育プログラム（リテラシーレベル）が具備すべき要素に「数理・データサイエンス・AI の基盤には統計学等の数学や情報科学が存在していること及びその役割を理解する内容が含まれている」（数理・データサイエンス・AI 教育プログラム認定制度会議，2020）ことを考慮すると、今後は文理を問わず数学の継続的な学習が必要になることが示唆される。

しかし、この目標である「文理を問わず」に関しては、慎重になる必要がある。なぜなら、文理で高等学校時に履修した数学に相違があるためである。渡辺他（2021）が「文系にするか、理系を選ぶかは数学によって決まるといわれている」と述べるように、文系学生は、数学Ⅲを含め履修していない科目や数学を学習していない期間が存在するため、理系学生と数学の基礎力に差があると考えられる。そのため、初級レベルの数理・データサイエンス・AI を習得には効果的な教育の必要性が示唆される。

よって、本研究の目的は、今後文系大学生に行われる「数理・AI・データサイエンスに関する科目」において効果的な教育を実施するために、実際該当する科目である「統計入門」で躓きや理解不足となっている内容について、テスト分析、特に誤答の分析を通して明らかにすることである。

## 2. 研究の方法

本研究の目標を達成するために、大学で開講されている統計入門のテストを分析した。回答時間は 60 分、有効回答数は 336 であった。なお、本研究は下関市立大学の倫理委員会（受付番号 2301-0417）の承認を得て行った。

学生の基礎力を適切に把握するため、実施したテストに対して、古典的テスト理論によって信頼性の確認を行った。その上で項目反応理論も加えてテストの項目分析を行い、最後に得られた答案について誤答の分析を行った。

## 3. 本研究のテスト項目及びテストの内容

テストの項目数は 36 で、尺度水準に関する項目が 8、時系列データ・折れ線グラフの用語に関する項目が 2、ジニ係数に関する項目が 1、平均・分散に関する項目が 4、加重平均に関する項目が 1、度

数分布に関する項目が6、標準得点に関する項目が2、変動係数に関する項目が1、2変量データに関する項目が3、正規分布に関する項目が3、2項分布に関する項目が4、区間推定に関する項目が1であった。項目の配列は表1の通りである。

なお、本研究における正答率の算出は完全正答のみを1とし、それ以外の誤答及び無答はすべて0とした。

## 4. 古典的テスト理論による分析およびその結果

### (1) 信頼性の確認

まずテストの信頼性の確認を行った。テストの信頼性を測る信頼性係数は直接求めることができないため、再検査法、平行検査法、折半法などを用いて推定することができるが、いずれも実施することが容易ではない。そのため、折半法で得られるすべての信頼性係数の推定値の平均値を信頼性係数の推定値とする内的整合性による方法が考えられる。しかし、すべての折半の組合せは、本テスト（36項目）の場合4537567650となり膨大である。そこで、熊谷他（2015）にある「弱平行測定を仮定して計算を簡便にした、クロンバックの$\alpha$係数を利用」した。算出したクロンバックの$\alpha$係数（Cronbach, 1951）は0.8306であり、熊谷他（2015）によると「学力などの能力を測定するテストでは0.80以上」を目安にするとあるため、本テストは信頼性があると判断した。

### (2) 各項目の識別力及び困難度の算出

古典テスト理論に基く各項目の識別力の指標として項目テスト間相関（IT相関)、項目リメインダ相関（IR相関)、困難度の指標として正答率を算出した結果は表1の通りである。テスト全体に対する各項目の寄与度を確認するため、当該項目を除外したクロンバックの$\alpha$係数（除外$\alpha$）を算出した結果も表1に含めた。

IT相関は、個々の項目と全項目の和との相関係数で、個々の項目の識別力を表す指標である。IT相関が低い場合はテストの信頼度を下げることになるため、テスト分析上は除外するべき項目となる。IT相関は当該項目を含んでいるため、個々の項目と残りの項目の和の相関係数を取ったものがIR相関である。熊谷他（2015）によると、IT相関及びIR相関の「目安は0.2以上」である。

表1のとおり、0.2以上の基準に満たないIT相関は項目3の比（率）尺度のみであるが、IR相関については項目2、3の比（率）尺度、項目12、27の平均の性質、項目28の分散の性質の5項目が該当した。また、除外$\alpha$係数が全体の$\alpha$係数である0.8306よりも大である項目、つまり除外することによりテストの信頼性が向上する項目は、項目2、3の比（率）尺度、項目27の平均の性質の3項目であった。

これらの結果から、項目2、3、12、27、28を削除し、算出した結果が表1の修正後の値である。修正後、IT相関及びIR相関を測り直したところ、0.2を下回る項目はなかった。

### (3) 各項目の識別力及び困難度の分析

表1より、IT相関が0.5以上の項目を識別力の高い項目と定義した場合、該当するのは項目18の標準得点、項目19の偏差値、項目20の変動係数、項目30と項目34の標準正規分布、項目32の2項分布の標準偏差、項目33の2項分布の確率であった。いずれの項目も複雑な計算はなく、定義を記憶することで正答に至るものであった。

正答率が0.4未満の項目を困難度の高い項目と定義した場合、該当するのは項目4の順序尺度、項目22の度数分布の中央値、項目35の2項分布の確率、項目36の区間推定の4項目であった。項目22、項目35及び項目36は複数の計算が必要となる項目で、項目4は尺度について正確な理解が必要なものであった。正答率が0.9以上の項目を困難度の低い項目と定義した場合、該当するのは項目5の名義尺度、項目14の階級値、項目29の標準正規分布で、いずれも単元の初期に学習する項目であった。

以上が古典的テスト理論による信頼性の確認、識別力、困難度が高いもしくは低い項目の抽出である。ただし、古典的テスト理論には、太田（2019）が「識別力や困難度の指標が受験者集団に依存する問題がある。つまり、正答率が低いことがわかったとしても、受験した生徒の能力が低いのか、問題が難しすぎるのかを区別するのが難しい」と述べる課題が存在する。

古典的テスト理論に対し項目反応理論は、加藤他

**表 1　項目の内容、困難度（正答率）、識別力等**

| | 項目の内容 | 困難度 | 識別力 | | 除外 $\alpha$ | 修正後の識別力 | | 修正後の除外 $\alpha$ |
|---|---|---|---|---|---|---|---|---|
| | | 正答率 | IT 相関 | IR 相関 | | IT 相関 | IR 相関 | |
| 1 | 比（率）尺度 1 | 0.8899 | 0.2528 | 0.2026 | 0.8293 | 0.2541 | 0.2007 | 0.8385 |
| 2 | 比（率）尺度 2 | 0.5655 | 0.2674 | 0.1877 | 0.8309 | 項目の削除 | | |
| 3 | 比（率）尺度 3 | 0.3988 | 0.0635 | −0.0189 | 0.8375 | 項目の削除 | | |
| 4 | 順序尺度 1 | 0.3839 | 0.4293 | 0.3591 | 0.8251 | 0.4387 | 0.3646 | 0.8343 |
| 5 | 名義尺度 | 0.9464 | 0.2892 | 0.2539 | 0.8286 | 0.2931 | 0.2557 | 0.8376 |
| 6 | 順序尺度 2 | 0.6190 | 0.4829 | 0.4166 | 0.8232 | 0.5042 | 0.4352 | 0.8319 |
| 7 | 間隔尺度 | 0.4613 | 0.3166 | 0.2383 | 0.8293 | 0.3258 | 0.2429 | 0.8387 |
| 8 | 順序尺度 3 | 0.5685 | 0.4253 | 0.3536 | 0.8253 | 0.4394 | 0.3638 | 0.8344 |
| 9 | 時系列データ | 0.7917 | 0.3338 | 0.2711 | 0.8278 | 0.3519 | 0.2860 | 0.8367 |
| 10 | 折れ線グラフ | 0.6875 | 0.3199 | 0.2475 | 0.8287 | 0.3254 | 0.2485 | 0.8381 |
| 11 | ジニ係数 | 0.8304 | 0.2879 | 0.2286 | 0.8289 | 0.2966 | 0.2337 | 0.8380 |
| 12 | 平均の性質 1 | 0.8304 | 0.2279 | 0.1669 | 0.8304 | 項目の削除 | | |
| 13 | 分散の性質 1 | 0.6161 | 0.3391 | 0.2638 | 0.8283 | 0.3114 | 0.2299 | 0.8390 |
| 14 | 階級値 | 0.9702 | 0.2456 | 0.2185 | 0.8294 | 0.2391 | 0.2101 | 0.8386 |
| 15 | 累積度数 | 0.7411 | 0.3863 | 0.3211 | 0.8264 | 0.4000 | 0.3313 | 0.8354 |
| 16 | 相対度数 | 0.5833 | 0.3967 | 0.3235 | 0.8264 | 0.3910 | 0.3126 | 0.8362 |
| 17 | 加重平均 | 0.8244 | 0.3309 | 0.2722 | 0.8277 | 0.3327 | 0.2702 | 0.8370 |
| 18 | 標準得点 | 0.4821 | 0.5382 | 0.4744 | 0.8211 | 0.5332 | 0.4647 | 0.8307 |
| 19 | 偏差値 | 0.4375 | 0.5649 | 0.5040 | 0.8201 | 0.5608 | 0.4954 | 0.8296 |
| 20 | 変動係数 | 0.4048 | 0.5396 | 0.4771 | 0.8211 | 0.5362 | 0.4692 | 0.8306 |
| 21 | 度数分布の平均値 | 0.6786 | 0.3449 | 0.2730 | 0.8279 | 0.3412 | 0.2644 | 0.8376 |
| 22 | 度数分布の中央値 | 0.0327 | 0.2382 | 0.2097 | 0.8295 | 0.2339 | 0.2035 | 0.8386 |
| 23 | 度数分布の最頻値 | 0.7262 | 0.3236 | 0.2542 | 0.8284 | 0.3425 | 0.2694 | 0.8373 |
| 24 | 2 変量データの分散 | 0.6964 | 0.4873 | 0.4250 | 0.8231 | 0.4899 | 0.4237 | 0.8323 |
| 25 | 共分散 | 0.5625 | 0.3800 | 0.3053 | 0.8270 | 0.4019 | 0.3237 | 0.8358 |
| 26 | 相関係数 | 0.5268 | 0.3887 | 0.3140 | 0.8267 | 0.3972 | 0.3182 | 0.8360 |
| 27 | 平均の性質 2 | 0.6280 | 0.2363 | 0.1576 | 0.8318 | 項目の削除 | | |
| 28 | 分散の性質 2 | 0.0804 | 0.2379 | 0.1941 | 0.8295 | 項目の削除 | | |
| 29 | 標準正規分布 1 | 0.9137 | 0.2911 | 0.2471 | 0.8285 | 0.2989 | 0.2522 | 0.8375 |
| 30 | 標準正規分布 2 | 0.5863 | 0.5303 | 0.4668 | 0.8215 | 0.5379 | 0.4709 | 0.8306 |
| 31 | 2 項分布の平均 | 0.8988 | 0.4357 | 0.3932 | 0.8254 | 0.4383 | 0.3932 | 0.8344 |
| 32 | 2 項分布の標準偏差 | 0.6935 | 0.6117 | 0.5596 | 0.8187 | 0.6332 | 0.5798 | 0.8271 |
| 33 | 2 項分布の確率 1 | 0.4464 | 0.5370 | 0.4735 | 0.8212 | 0.5585 | 0.4927 | 0.8297 |
| 34 | 標準正規分布 3 | 0.7113 | 0.5406 | 0.4833 | 0.8213 | 0.5386 | 0.4772 | 0.8306 |
| 35 | 2 項分布の確率 2 | 0.2917 | 0.4848 | 0.4231 | 0.8232 | 0.4899 | 0.4245 | 0.8323 |
| 36 | 区間推定 | 0.3363 | 0.4636 | 0.3978 | 0.8239 | 0.4634 | 0.3932 | 0.8333 |

(2014) が「テストに含まれる項目の難易度とそのテストの受験者の能力を分離して表現できる」と述べる利点がある。そのため、テストの受験者集団に依存しない識別力と困難度を推定するために、項目反応理論による分析を行った。

## 5. 項目反応理論適用の条件の確認・分析結果

### (1) 適応する項目反応曲線及びその条件

本テスト分析では、受験者集団の能力値、正答確率、識別力及び困難度を図1の通りに設定し、2パラメータ・ロジスティック・モデルを用いて分析を行った。

能力値$\theta$の受験者が、項目$j$に正答する確率を$Pj(\theta)$、識別力を$a_j$、困難度を$b_j$とする。このとき、$Pj(\theta)$は、次の式で与えられる。

$$Pj(\theta)=\frac{1}{1+\exp\{-a_j(\theta-b_j)\}}$$

**図1　項目反応理論の条件設定**

項目反応理論は、受験者集団に依存しない識別力と困難度を推定できる利点がある一方、適用する条件として一次元性及び局所独立性の成立が必要となる。一次元性は、能力値が1つの指標でそれぞれの項目の正答確率の分布が得られていることを保証するもので、局所独立性は、ある項目の正答が、他項目の正答と独立であるか、影響を与えていないかを保証するものである。項目反応理論で分析するために、それぞれの確認を行った。なお、古典的テスト理論によって識別力が低かった項目2、3、12、27、28を削除したのち、一次元性、局所独立性を測定したところ、局所独立性について後に示す$Q_3$統計量で大きいペアが存在したため、原因となる項目6、18、19、32を削除した27項目で検討した。

### (2) 一次元性の確認

因子数の決定には、固有値を降順にプロットし、推移がなだらかになる直前までを因子数とするスクリープロット基準で検証した。27項目の固有値を降順にプロットした結果は図2の通りで、第2固有値以降の減衰状況から一次元性は成立するものと判断した。固有値の算出には、熊谷（2009）を用いた。

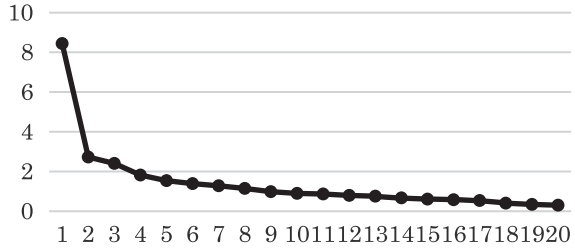


**図2　固有値を降順に並べたスクリープロット**
**（縦軸は固有値，横軸は固有値番号）**

### (3) 局所独立性の確認

局所独立性は、$Q_3$統計量（Yen，1984）で検討した。本テストの全36項目から、IT相関、IR相関が0.2よりも低い5項目を除いた31項目で検討した際に、前述のとおり$Q_3$統計量の絶対値が0.2を大きく超えたペアの項目が存在した。そのため、その原因となる項目6、18、19、32を削除し、測りなおした結果が表2である。異なる2個の項目のペア351のうち9つのペアについては、$Q_3$統計量が0.2と0.3の間の値となった。具体的には、項目1の比（率）尺度と項目5の名義尺度、項目4、8の順序尺度、項目8の順序尺度と項目29の標準正規分布、項目14の階級値と項目31の2項分布の平均、項目15の累積度数と項目16の相対度数、項目15の累積度数と項目26の相関係数、項目20の変動係数と項目35の2項分布の確率、項目21の度数分布の平均値と項目23の度数分布の最頻値、項目25の共分散と項目26の相関係数であった。これらの項目が能力値、識別力及び困難度に与える影響の検証については、今後の課題としたい。

### (4) 項目反応理論による分析結果

項目反応理論による識別力、困難度の推定結果は表3の通りであり、推定には熊谷（2009）を用いた。

表3から困難度の高い項目を$b_j$の値が0.5以上のものと定義した場合、項目4の順序尺度、項目22の度数分布の中央値、項目35の2項分布の確率、項目36の区間推定の4項目が該当したが、古典的テスト理論で行った結果と変化は認められなかった。

また、表3から識別力の高い項目を$a_j$の値が1.5以上のものと定義した場合、項目22の度数分布の中央値、項目30及び34の標準正規分布、項目31の2項分布の平均、項目33及び35の2項分布の確率の6項目が該当したが、項目22は古典的テスト理論の結果と大きく異なった。

次に困難度及び識別力を順序尺度と捉えて序列にしたものが表4である。

困難度は、項目1、10、21、30、34について古典

表 2　$Q_3$ 統計量

| 項目 | 1 | 4 | 5 | 7 | 8 | 9 | 10 | 11 | 13 | 14 | 15 | 16 | 17 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 29 | 30 | 31 | 33 | 34 | 35 | 36 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | 0.02 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 5 | 0.22 | 0.05 | 1 | | | | | | | | | | | | | | | | | | | | | | | | |
| 7 | 0.06 | 0.08 | 0.02 | 1 | | | | | | | | | | | | | | | | | | | | | | | |
| 8 | 0.02 | 0.29 | 0.06 | 0.17 | 1 | | | | | | | | | | | | | | | | | | | | | | |
| 9 | 0 | 0.06 | 0.12 | 0.06 | 0.06 | 1 | | | | | | | | | | | | | | | | | | | | | |
| 10 | 0.07 | 0.03 | 0.05 | 0.06 | 0 | 0.08 | 1 | | | | | | | | | | | | | | | | | | | | |
| 11 | 0.07 | 0.08 | 0.02 | 0.09 | 0.04 | 0.01 | 0.09 | 1 | | | | | | | | | | | | | | | | | | | |
| 13 | 0.04 | 0.04 | 0.04 | 0 | 0.05 | 0.03 | 0.03 | 0.02 | 1 | | | | | | | | | | | | | | | | | | |
| 14 | 0.1 | 0.04 | 0.06 | 0.04 | 0.08 | 0.07 | 0.05 | 0.04 | 0.01 | 1 | | | | | | | | | | | | | | | | | |
| 15 | 0.17 | 0.03 | 0.06 | 0.02 | 0.02 | 0.07 | 0.03 | 0.19 | 0.03 | 0.2 | 1 | | | | | | | | | | | | | | | | |
| 16 | 0.07 | 0.08 | 0.06 | 0.02 | 0.02 | 0.1 | 0.01 | 0.2 | 0.08 | 0.09 | 0.25 | 1 | | | | | | | | | | | | | | | |
| 17 | 0.04 | 0 | 0.04 | 0.14 | 0.15 | 0.04 | 0.02 | 0.16 | 0.05 | 0 | 0.12 | 0.04 | 1 | | | | | | | | | | | | | | |
| 20 | 0.05 | 0.02 | 0.07 | 0.06 | 0.02 | 0.1 | 0.01 | 0.02 | 0 | 0.02 | 0.1 | 0.03 | 0.02 | 1 | | | | | | | | | | | | | |
| 21 | 0.01 | 0.01 | 0.06 | 0.08 | 0.02 | 0 | 0.04 | 0.13 | 0.07 | 0.03 | 0.02 | 0.03 | 0.02 | 0.06 | 1 | | | | | | | | | | | | |
| 22 | 0.03 | 0.05 | 0.01 | 0.04 | 0.04 | 0.03 | 0.06 | 0.03 | 0.02 | 0.01 | 0.02 | 0.01 | 0.02 | 0.01 | 0 | 1 | | | | | | | | | | | |
| 23 | 0.03 | 0.04 | 0.07 | 0.04 | 0.04 | 0.05 | 0.14 | 0.04 | 0.05 | 0.11 | 0.08 | 0.03 | 0.06 | 0.06 | 0.24 | 0.06 | 1 | | | | | | | | | | |
| 24 | 0.06 | 0.03 | 0.08 | 0 | 0.04 | 0.12 | 0.05 | 0.1 | 0.07 | 0.03 | 0.12 | 0.04 | 0.12 | 0.11 | 0.02 | 0.03 | 0.09 | 1 | | | | | | | | | |
| 25 | 0.06 | 0.09 | 0.02 | 0.06 | 0.04 | 0.01 | 0.06 | 0.07 | 0.08 | 0.09 | 0.13 | 0.06 | 0.04 | 0 | 0.02 | 0.13 | 0.03 | 0.03 | 1 | | | | | | | | |
| 26 | 0.02 | 0.03 | 0.05 | 0.07 | 0.11 | 0.08 | 0.06 | 0.16 | 0.15 | 0 | 0.23 | 0.1 | 0.09 | 0.06 | 0.02 | 0.11 | 0.06 | 0 | 0.23 | 1 | | | | | | | |
| 29 | 0.02 | 0.07 | 0.09 | 0.07 | 0.22 | 0.03 | 0.03 | 0.02 | 0.04 | 0.1 | 0.18 | 0.14 | 0.02 | 0.07 | 0.05 | 0.03 | 0.04 | 0.06 | 0.1 | 0.04 | 1 | | | | | | |
| 30 | 0.01 | 0.03 | 0.01 | 0.07 | 0.12 | 0.02 | 0.03 | 0.01 | 0.08 | 0.12 | 0.18 | 0.15 | 0.09 | 0.04 | 0.03 | 0.04 | 0 | 0.04 | 0.07 | 0 | 0.07 | 1 | | | | | |
| 31 | 0.02 | 0.13 | 0.2 | 0.05 | 0.07 | 0.08 | 0.14 | 0.07 | 0.08 | 0.25 | 0.03 | 0 | 0.14 | 0.04 | 0.15 | 0.08 | 0.05 | 0.07 | 0.18 | 0.11 | 0.06 | 0.1 | 1 | | | | |
| 33 | 0.08 | 0.14 | 0.03 | 0.12 | 0.11 | 0.08 | 0.12 | 0.07 | 0.13 | 0.01 | 0.01 | 0.03 | 0.12 | 0.06 | 0.2 | 0.1 | 0.17 | 0.1 | 0.14 | 0.12 | 0.01 | 0.11 | 0.04 | 1 | | | |
| 34 | 0.07 | 0.18 | 0.09 | 0.07 | 0.1 | 0.12 | 0 | 0.11 | 0.04 | 0.1 | 0.03 | 0.01 | 0.12 | 0.16 | 0.17 | 0.04 | 0.13 | 0.19 | 0.14 | 0.03 | 0.05 | 0.16 | 0.12 | 0.01 | 1 | | |
| 35 | 0.04 | 0.12 | 0.08 | 0.06 | 0.08 | 0.16 | 0.1 | 0.03 | 0.05 | 0.03 | 0.13 | 0.17 | 0.01 | 0.22 | 0.05 | 0.14 | 0.04 | 0.03 | 0 | 0.12 | 0.04 | 0.09 | 0.01 | 0.01 | 0.03 | 1 | |
| 36 | 0.02 | 0.14 | 0.04 | 0 | 0.02 | 0.09 | 0.07 | 0.09 | 0 | 0.05 | 0 | 0 | 0.07 | 0.12 | 0.01 | 0.05 | 0.06 | 0.07 | 0.12 | 0.03 | 0.06 | 0.1 | 0.02 | 0.12 | 0.1 | 0.01 | 1 |

的テスト理論による結果と項目反応理論による結果で差があったものの、大きな差ではなかった。

識別力は、項目 4、5、8、14、22、25、29、31 について、古典的テスト理論による結果と項目反応理論による結果で大きな差があった。これから、標本依存性、項目依存性の可能性も配慮して分析を行う。

**表 3　項目反応理論による識別力・困難度の推定**

| 項目 | 古典的テスト理論 CTT | | | 項目反応理論 IRT | |
|---|---|---|---|---|---|
| | 困難度 | 識別力 | | 困難度 | 識別力 |
| | 正答率 | IT 相関 | IR 相関 | $b_j$ | $a_j$ |
| 1 | 0.8899 | 0.2711 | 0.2063 | −3.2413 | 0.7013 |
| 4 | 0.3839 | 0.4253 | 0.3328 | 0.6737 | 0.8016 |
| 5 | 0.9464 | 0.3135 | 0.2683 | −2.7066 | 1.3276 |
| 7 | 0.4613 | 0.3346 | 0.2332 | 0.3358 | 0.4889 |
| 8 | 0.5685 | 0.4141 | 0.3189 | −0.4091 | 0.7565 |
| 9 | 0.7917 | 0.3805 | 0.3012 | −1.7191 | 0.8988 |
| 10 | 0.6875 | 0.3429 | 0.2494 | −1.5218 | 0.5535 |
| 12 | 0.8304 | 0.2989 | 0.2219 | −2.5863 | 0.6679 |
| 13 | 0.6161 | 0.3289 | 0.2298 | −0.9403 | 0.5358 |
| 14 | 0.9702 | 0.2519 | 0.2168 | −3.2739 | 1.2936 |
| 15 | 0.7411 | 0.3954 | 0.3105 | −1.3544 | 0.9039 |
| 16 | 0.5833 | 0.4088 | 0.3136 | −0.4731 | 0.8108 |
| 17 | 0.8244 | 0.3633 | 0.2883 | −1.9455 | 0.9224 |
| 20 | 0.4048 | 0.4666 | 0.3767 | 0.4551 | 1.0400 |
| 21 | 0.6786 | 0.3671 | 0.2744 | −1.3872 | 0.5788 |
| 22 | 0.0327 | 0.2433 | 0.2064 | 2.6139 | 1.7981 |
| 23 | 0.7262 | 0.3689 | 0.2807 | −1.6611 | 0.6387 |
| 24 | 0.6964 | 0.4847 | 0.4025 | −0.8727 | 1.2187 |
| 25 | 0.5625 | 0.4268 | 0.3325 | −0.3642 | 0.7795 |
| 26 | 0.5268 | 0.4115 | 0.3152 | −0.1503 | 0.8076 |
| 29 | 0.9137 | 0.3118 | 0.2550 | −2.3892 | 1.2189 |
| 30 | 0.5863 | 0.5472 | 0.4653 | −0.3222 | 1.5247 |
| 31 | 0.8988 | 0.4247 | 0.3686 | −1.674 | 2.0778 |
| 33 | 0.4464 | 0.5539 | 0.4719 | 0.1832 | 1.9163 |
| 34 | 0.7113 | 0.5397 | 0.464 | −0.7598 | 1.8283 |
| 35 | 0.2917 | 0.5125 | 0.4339 | 0.7655 | 1.7642 |
| 36 | 0.3363 | 0.4895 | 0.4053 | 0.6691 | 1.3782 |

**表 4　困難度、識別力が高い項目の序列**
**表内の数値は項目番号**

| 順位 | 困難度 | | 識別力 | | |
|---|---|---|---|---|---|
| | CTT | IRT | CTT | | IRT |
| | 正答率 | $b_j$ | IT 相関 | IR 相関 | $a_j$ |
| 1 | 22 | 22 | 33 | 33 | 31 |
| 2 | 35 | 35 | 30 | 30 | 33 |
| 3 | 36 | 4 | 34 | 34 | 34 |
| 4 | 4 | 36 | 35 | 35 | 22 |
| 5 | 20 | 20 | 36 | 36 | 35 |
| 6 | 33 | 7 | 24 | 24 | 30 |
| 7 | 7 | 33 | 20 | 20 | 36 |
| 8 | 26 | 26 | 25 | 31 | 5 |
| 9 | 25 | 30 | 4 | 4 | 14 |
| 10 | 8 | 25 | 31 | 25 | 29 |
| 11 | 16 | 8 | 8 | 8 | 24 |
| 12 | 30 | 16 | 26 | 26 | 20 |
| 13 | 13 | 34 | 16 | 16 | 17 |
| 14 | 21 | 24 | 15 | 15 | 15 |
| 15 | 10 | 13 | 9 | 9 | 9 |
| 16 | 24 | 15 | 23 | 17 | 16 |
| 17 | 34 | 21 | 21 | 23 | 26 |
| 18 | 23 | 10 | 17 | 21 | 4 |
| 19 | 15 | 23 | 10 | 5 | 25 |
| 20 | 9 | 31 | 7 | 29 | 8 |
| 21 | 17 | 9 | 13 | 10 | 1 |
| 22 | 12 | 17 | 5 | 7 | 12 |
| 23 | 1 | 29 | 29 | 13 | 23 |
| 24 | 31 | 12 | 12 | 12 | 21 |
| 25 | 29 | 5 | 1 | 14 | 10 |
| 26 | 5 | 1 | 14 | 22 | 13 |
| 27 | 14 | 14 | 22 | 1 | 7 |

## 6. 項目の誤答分析

項目1から項目8は、Stevens（1951）が提唱した4水準に関するものである。この4水準を学ぶ理由は、谷口（2017）が「構成する尺度がどの水準にあたるかは、その後の統計処理や研究する心理的概念に深くかかわることである」と指摘する通りである。構成する尺度水準によって可能な計算や適したグラフが異なるため、これを誤ると、間違ったデータを提示することやデータをわかりやすくするためのグラフがわかりにくくなることもある。

4水準すべて正答した学生は表5の通り、17人で全体の5.06%に留まり、全体の94.94%は少なくとも1項目は誤答をしていた。この結果から、統計処理を行う上で土台となる4水準の理解が十分ではないことが示唆された。4水準を量的データの正誤、質的データの正誤に限定した結果も表5の通りであるが、いずれの理解も十分ではないことが示唆された。この結果から4水準の理解を徹底させる必要性がある。以後、4水準の理解度を名義尺度、順序尺度、間隔尺度、比（率）尺度の順に分析していく。

**表5　4水準、量的データ、質的データの正答及び正答率**

| | 正答（率） | 不正答（率） |
|---|---|---|
| 4水準 | 17（5.06） | 319（94.94） |
| 量的データ | 34（10.12） | 302（89.88） |
| 質的データ | 91（27.08） | 245（72.92） |

### (1) 名義尺度

名義尺度に関する項目は項目5である。病因を名義尺度と回答する項目で、結果は表6の通りであった。学生の94.64%が正答であり、誤答となった学生及び無答の学生を合わせると5.36%であった。

**表6　名義尺度に関する項目5の正答及び正答率**

| 項目 | 正答（率） | 誤答（率） | 無答（率） |
|---|---|---|---|
| 5 | 318（94.64） | 16（4.76） | 2（0.60） |

本項目が正答できていない学生は、他の項目も多くが誤答もしくは無答となっており、4水準の理解が得られていないことが示唆される。4水準の理解は統計処理を行う際に重要であり、理解不足が躓きの原因にもつながるため、具体例などを通して理解させる必要がある。

誤答の詳細は、表7の通りである。過半数を超える9名が存在しない尺度水準を回答しており、4水準の定義を把握していないことが分かる。その他の7人中、6人が量的データである間隔尺度及び比（率）尺度を回答しており、質的データである順序尺度を回答した学生は1人であった。この結果から、量的データ及び質的データの理解が得られていないことが示唆される。4水準の理解をさせるために、4水準の定義を正確に記憶させることと合わせて量的データ及び質的データの正確な理解が必要になると考える。

**表7　名義尺度に関する項目5の誤答例**

| 人数 | 誤答内容 |
|---|---|
| 4 | 間隔 |
| 2 | 比（率）、相対、質 |
| 1 | 順序、記号、質量、差異、質的、心的 |

なお、本項目は94.64%の学生が正答となったが、本項目を回答できることと、それがどのような統計処理に応用されるのかを理解しているのかは異なる。そのため、谷口（2017）が指摘するように名義尺度は「各カテゴリに分類される度数（人数など）を用いた統計は可能であり、全体に対する割合の算出、最頻値、クロス集計、$\chi^2$ 検定などの度数に関するノンパラメトリック検定などが可能である。」こと、つまり度数を用いることで様々な分析ができることを教育する必要があると考える。

### (2) 順序尺度

順序尺度に関する項目は、4、6、8である。項目4は震度を、項目6は相撲番付（横綱、大関、関脇など）を、項目8は授業のやり方の評価（非常に良い、中程度、常に悪い）を順序尺度と回答するもので、結果は表8の通りであった。前項目の名義尺度と比較して、正答率は大きく減少した。

**表8　順序尺度に関する項目4、6、8の正答及び正答率**

| 項目 | 正答（率） | 誤答（率） | 無答（率） |
|---|---|---|---|
| 4 | 129（38.39） | 201（59.82） | 6（1.79） |
| 6 | 208（61.90） | 123（36.61） | 5（1.49） |
| 8 | 191（56.85） | 140（41.67） | 5（1.49） |

順序尺度については、数量化されている震度と数量化されていない相撲番付及び授業のやり方の評価

で誤答の型が分かれた。

数量化された震度の尺度に関する誤答の詳細は、表 9 の通りで比（率）尺度及び間隔尺度と間違える学生が多かった。それらの合計である 171 人は、正答者数の 129 人を上回る数であるため、震度が間隔尺度及び比（率）尺度とはならない、つまり量的データとはならない理由を説明する必要がある。

前述を踏まえ順序尺度は、谷口（2017）が指摘する「昇順・降順による順序比較、代表値としての中央値や、順位相関係数を求めることができるが、順位の和や差を求めることができない」ことを教育する必要があると考える。

**表 9　順序尺度に関する項目 4 の誤答例**

| 人数 | 誤答内容 |
|---|---|
| 107 | 比（率） |
| 64 | 間隔 |
| 18 | 名義 |
| 5 | 量 |
| 1 | 頻分、震度、質、順位、序列、等級、順番 |

数量化されていない相撲番付の尺度及び授業のやり方の評価の尺度について誤答の詳細は表 10 であり、名義尺度と間違える学生が過半数を超えていた。

**表 10　順序尺度に関する項目 6、8 の誤答例**

| 項目 6 | | 項目 8 | |
|---|---|---|---|
| 人数 | 誤答内容 | 人数 | 誤答内容 |
| 93 | 名義 | 79 | 名義 |
| 15 | 間隔 | 22 | 間隔、比（率） |
| 3 | 質 | 6 | 質 |
| 2 | 比（率）、順位 | 2 | 量、順位 |
| 1 | 度数、相対、量、価値、序列、順位、等級、順番 | 1 | 程度、心的、価値、序列、順列、等級、順番 |

この結果から多くの学生は、対象が数量化されていない場合は名義尺度と回答する確率が高いことが示唆される。

そのため名義尺度にはない順序の概念について説明する必要がある。順序については、震度の例のようにダミー変数を用いることで数量化して説明する方法もある。

ただし、相撲番付（横綱、大関、関脇など）及び授業のやり方の評価（非常に良い、中程度、常に悪い）に対してダミー変数を付与して数量化する場合は、前述の通り名義尺度と回答する学生は減少するが、比（率）尺度や間隔尺度と回答する学生が増加する可能性が示唆される。特に、GPA のように、秀に 4、優に 3、良に 2、可に 1 を付与して、数値間に等間隔性が保証されていないものを等間隔と仮定して考えるものもあり誤解を生みやすい。また、井上（2015）が述べる通り「統計学の教科書は、例えば算術平均値は間隔尺度以上でしか計算できないと指摘するが、多くの統計学的分析では順序尺度を間隔尺度とみなしている」のが現状である。この点については、注意事項として慎重に扱い教育する必要がある。

### (3) 間隔尺度

間隔尺度に関する項目は項目 7 である。テストの点数を間隔尺度と回答する項目で、結果は表 11 の通りであった。

**表 11　間隔尺度に関する項目 7 の正答及び正答率**

| 項目 | 正答（率） | 誤答（率） | 無答（率） |
|---|---|---|---|
| 7 | 155（46.13） | 176（52.38） | 5（1.49） |

誤答の詳細は、表 12 の通りであるが、比（率）尺度の誤答が多くを占める結果となり、誤解しやすいことが分かった。

**表 12　間隔尺度に関する項目 7 の誤答例**

| 人数 | 誤答内容 |
|---|---|
| 141 | 比（率） |
| 12 | 順序 |
| 9 | 量 |
| 8 | 名義 |
| 2 | 質 |
| 1 | 順列、得点、絶対、相対 |

間隔尺度と比（率）尺度の違いはゼロを基点とした絶対原点の有無である。間隔尺度は絶対原点がなく、比（率）尺度は絶対原点が存在する。

項目 7 のテストの点数は間隔尺度であるが、それはテストの点数が 0 であった場合、そのテストに関する学力が全くないことを意味しない。例えば、大学生に算数オリンピックの問題を出題した場合を考える。大学生は、義務教育を卒業しているため算数を履修し、算数の知識はあるが、算数オリンピックの問題が解けず 0 点になる可能性は考えられる。

しかし、このときの0点は算数オリンピックの問題が解けなかっただけであり、算数の能力が0であることを意味しない。テストの点数は比（率）尺度と混同しやすいが、具体例を通して説明することで正確な理解につなげることができると考える。

### (4) 比（率）尺度

比（率）尺度に関する項目は、1、2、3である。項目1は身長、項目2は時間の経過、項目3は日較差を比（率）尺度と回答するもので、結果は表13の通りであった。項目1から3で、正答率が減少するように作為して作問したため、結果は予想の範囲内に概ね収まった。なお、項目1から項目3まですべて正答であった学生は71人で全体の21.13%、少なくとも1つの項目は不正答であった学生は265人で全体の78.87%を占めていた。

**表13　比(率)尺度に関する項目1、2、3の正答及び正答率**

| 項目 | 正答（率） | 誤答（率） | 無答（率） |
|---|---|---|---|
| 1 | 299（88.99） | 34（10.12） | 3（0.89） |
| 2 | 190（56.55） | 142（42.26） | 4（1.19） |
| 3 | 134（39.88） | 194（57.74） | 8（2.38） |

項目1の正答率は良好であった。身長は体重と合わせて比（率）尺度の具体例として、最初に挙げられることが多いため、良好な結果になったと考える。項目1の誤答の詳細は表14の通りであった。

**表14　比（率）尺度に関する項目1の誤答例**

| 人数 | 誤答内容 |
|---|---|
| 19 | 間隔 |
| 7 | 名義 |
| 2 | 順序、量 |
| 1 | 相対、丈、序列、度数 |

項目2の誤答の詳細は表15の通りで、誤答の92.96%は間隔尺度と回答していた。時間が間隔尺度であるため、時間の経過も同様に考え誤答になった可能性が示唆される。

そのため、時間と時間の経過の相異点を、絶対原点の有無を通して、具体的に明示する必要があると考える。

**表15　比（率）尺度に関する項目2の誤答例**

| 人数 | 誤答内容 |
|---|---|
| 132 | 間隔 |
| 6 | 順序 |
| 2 | 量 |
| 1 | 名義、相対 |

項目3の誤答の詳細は表16の通りで、誤答の87.63%は、間隔尺度と回答していた。項目3が項目2よりも誤答者が多い理由は、日較差の定義である「最高気温と最低気温の差」にあると考える。定義に差があるため、差に着目する尺度である間隔尺度を回答とした可能性が示唆される。そのため定義を記憶するのみならず意味も理解する必要がある。

**表16　比（率）尺度に関する項目3の誤答例**

| 人数 | 誤答内容 |
|---|---|
| 170 | 間隔 |
| 6 | 順序、名義 |
| 5 | 量 |
| 2 | 質 |
| 1 | 気温、絶対、格差、相対、重 |

なお項目3は、IT相関が0.0635、IR相関が-0.0189と低いため、テストの項目としては熊谷他(2015)の通り「今回測定している力をまったく反映していない」ことになる。

しかし、テストの信頼度を向上させる項目としては不適であっても、比（率）尺度を理解させる項目としては重要であり、効果的に理解させるためにも誤答の分析は必要である。

そして、この項目のIT相関及びIR相関の値から、統計のテストにおいて点数が高い学生も、点数が低い学生も一様に比（率）尺度を間隔尺度と間違えることがわかる。つまり、間隔尺度と比（率）尺度の違いについては理解不足の学生が多いことが示唆される。そのため、両尺度については、具体例を多く提示することによって理解を促進させる必要性がある。

### (5) 加重平均

加重平均に関する項目は項目17である。項目17は、表17の条件に対して、それぞれの点数を英語：数学：国語：社会＝3：4：1：2と傾斜配点した場

合の加重平均を求める項目であった。

**表 17　項目 17 の条件設定**

| 英語 | 数学 | 国語 | 社会 |
|---|---|---|---|
| 70 | 50 | 60 | 80 |

平均の理解度は、良好であり、計算間違いを除けば全員が正答になることを授業時に確認したため、加重平均の理解度を問うた。結果は表 18 の通りで、無答の学生が 1 人存在した。

**表 18　加重平均に関する項目 17 の正答及び正答率**

| 項目 | 正答（率） | 誤答（率） | 無答（率） |
|---|---|---|---|
| 17 | 277（82.44） | 58（17.26） | 1（0.29） |

正答は図 3 の通りである。

$$\frac{70 \times 3 + 50 \times 4 + 60 \times 1 + 80 \times 2}{10} = 63$$

**図 3　項目 17 の正答**

誤答の詳細は表 19 の通りで、一番多かった誤答及び次点の誤答は加重平均の定義を明確に理解していないものであった。

**表 19　加重平均に関する項目 17 の誤答例**

| 人数 | 誤答内容 |
|---|---|
| 28 | 630/4 ＝ 157.5（計算間違い 6 名を含む） |
| 8 | 63/4（計算間違い 1 名を含む） |
| 4 | 65（計算間違い） |
| 3 | 66（計算間違い） |
| 2 | 26、68（計算間違い） |
| 1 | 53、57、62、64、67、70、73、77.5、83 |
| 2 | その他 |

一番多かった誤答を具体的に見ると、分母を 10 とするのではなく 4 としたものが 28 人で誤答の中の 48.57%を占めていた。なおこの 28 人の中の 6 人は計算間違いもしており、具体的には 90 と間違えた学生が 2 人、187.5、137.5、132.5、12.6 と間違えた学生がそれぞれ 1 人存在していた。

誤答の次点は、先ほどの数値を更に 10 で割ったものであった。

これらの誤答から、平均値に関する理解の不足が示唆される。平均値を求める計算は、小学校で学習するが、本結果から形式的な計算の方法の理解に留まっている可能性がある。

なぜなら、平均値は最大値を超えることも最小値より小さくなることもないからである。本項目で考察すると、表 17 にある 4 科目の平均値が最大値 80 を超えることも、最小値である 50 を下回ることもない。しかし、一番多い誤答の 630/4 つまり 157.5 は、最大値の 80 を超えており、次点の誤答である 15.75 は最小値である 50 を下回っている。これらは、平均値に関する正しい理解があれば、間違いであることは容易にわかる。その後は平均値の定義に従って修正することとなるが、平均値に関する正しい理解がなければ、修正する必要があることがわかっても、どのように修正するのかがわからず、その結果間違ったまま回答を提出することになる。そのため、平均値を計算して求めることができるのみならず、その計算は何を意味しているのか理解する必要がある。特に、文部科学省（2019）にある通り「数学的活動を通して、その有用性を認識」させない限り、意味の理解が形式的なものや抽象的なものに留まり、具体的に応用できない可能性がある。平均値の意味を理解させたうえで、具体的な問題を通して理解させ有用性につなげる必要があると考える。

### (6) 標準得点

標準得点に関する項目は項目 18 である。項目 18 は、表 20 から標準得点に求める問題である。標準得点の定義を正確に記憶しているかを把握するため、表 20 で与えた条件を標準偏差ではなく分散とした。つまり、曖昧に定義を記憶している学生が、偶然正解することがないように、数値を調整して出題した。

**表 20　項目 18 の条件設定**

| 得点 | 平均点 | 分散 |
|---|---|---|
| 58 | 50 | 16 |

正答は図 4 の通りであり、結果は表 21 の通りであった。

$$\frac{58 - 50}{\sqrt{16}} = 2$$

**図 4　項目 18 の正答**

表 21　標準得点に関する項目 18 の正答及び正答率

| 項目 | 正答（率） | 誤答（率） | 無答（率） |
|---|---|---|---|
| 18 | 162（48.21） | 157（46.73） | 17（5.06） |

誤答の詳細は表 22 の通りで、点数から平均点を引き、分散で割った回答が 106 人と一番多く、誤答の中で 67.52%を占めており想定していた通りの結果となった。誤答の次点は、標準偏差を回答したものあった。この誤答については問題文を正確に読んでいない可能性と標準得点の定義を正確に記憶していない可能性が示唆される。

他は、計算過程がなく 1 と回答したもの、変動係数を回答したものが続いた。標準得点は、偏差値や知能指数などの統計量を求める場合などを含め、様々な統計処理で活用される。誤答例の通り、計算間違いを誘発するような難しい定義ではないため、理解させ記憶させる必要がある。

表 22　標準得点に関する項目 18 の誤答例

| 人数 | 誤答内容 |
|---|---|
| 106 | 0.5 （点数－平均）／分散 |
| 11 | 4　標準偏差を回答 |
| 9 | 1 |
| 5 | 0.08　変動係数を回答 |
| 4 | 54、0.8　変動係数を回答し、計算ミス |
| 3 | 8　変動係数を%で回答 |
| 2 | 55、60 |
| 1 | 0.1、0.16、0.2、0.32、0.84、1.16、5、34、50、52、54.9 |

## 7. まとめ

本稿の目的は、統計入門に関するテストを分析することで、学生の理解不足及び躓き箇所を明らかにし、効果的な教育につなげることであった。

本研究結果から、Stevens の 4 水準について正確に理解している学生が少ないことがわかった。具体的には、名義尺度の理解はあるものの、順序尺度、間隔尺度、比（率）尺度については、理解不足が顕著に表れた。

Stevens の 4 水準は高等学校まででは学習せず、大学で学習する学生が多い。今後の統計処理に基礎となる重要な概念であることから、本研究結果を用いて、理解不足になりやすい点、誤答になりやすい点を考慮した教育が必要になると考える。

加重平均の項目を通して、平均値の機械的な計算はできるものの、意味の理解に不足が見られる学生が存在していた。

また標準得点については定義を覚えていない学生、定義の記憶が曖昧な学生が散見された。

そのため、本研究で得られた誤答例を通して曖昧な理解や曖昧な記憶にならないように、定着を図る教育を実施する必要がある。

**引用・参考文献**


Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests., Psychometrika, 16 (3), 297-334. https://www.doi.org/10.1007/BF02310555

井上信次（2015）．項目反応理論に基づく順序尺度の等間隔性―質問紙調査の回答選択肢（3 ～ 5 件法）の等間隔性と回答のしやすさ―．川崎医療福祉学会誌，25 (1), 23-35

加藤健太郎・山田剛史・川端一光（2014）．R による項目反応理論．オーム社.

熊谷龍一（2009）．初学者向けの項目反応理論分析プログラム EasyEstimation シリーズの開発．日本テスト学会誌，5，107-118.

熊谷龍一・荘島宏二郎（2015）．教育心理学のための統計学．誠信書房.

文部科学省（2019）．高等学校学習指導要領（平成 30 年告知）解説数学編理数編．学校図書.

内閣府（2022）．統合イノベーション戦略推進会議 AI 戦略 2022, https://www8.cao.go.jp/cstp/ai/aistrategy2022_honbun. pdf（2023.10.5 参照）

太田絵梨子（2021）．数学の概念的理解を評価するテストの提案と実践的検討．教育心理学研究，69，204-220.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), Handbook of experimental psychology，1-49.

谷口高士（2017）．心理評価実験における尺度構成の方法．日本音響学会誌，73（12），774-782

渡辺信・青木孝子（2021）．高大接続：大学文系学部入試数学必須の動き．日本科学教育学会研究会研究報告，35（8），23-26.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement,. 8，125-145