データサイエンスリテラシーレベル教育のための データ処理実演ツールについて

鈴木 治郎 [*1]

[*1] 信州大学 全学教育機構

[*1]szkjiro@shinshu-u.ac.jp

キーワード　データサイエンス教育，再現性，CUI

1　はじめに

データサイエンティスト教育の充実が国際的に求められている今日，国内では文部科学省が AI 戦略 2019 を発表した [1]．そこでは，2024 年に 50 万人の高等教育在学生がデータサイエンス（以下，DS と略す）リテラシーレベルを履修，さらに 2025 年には内半数近くが応用基礎レベルの履修を始めることが望まれている．そうした高等教育を取り巻く状況の中，各大学等ではその準備を進めている最中にある．この DS リテラシーレベルに対して，文部科学省ではそのカリキュラムの認定制度を始めており，本年度（2021 年度）には合わせて 89 校の大学・高等専門学校が認定を受けている．

この DS リテラシーレベルカリキュラムの策定にあたっては「実データ（または模擬データ）を用いた講義を行うことが望ましい」と書かれており，たとえば放送大学で放送の始まった「データサイエンス基礎から応用」のコース（上記リテラシーレベルの内「基礎」に相当）では，京都大学作成の講座（45 分全 8 回）を放送している．そこで実データの実演に用いられるツールには表計算ソフトの Excel が採用されている．

データサイエンスなどのデータ分析目的に Excel は広く使われていることから，入門期教育のためのこの選択は一見妥当に思えるが，その妥当性はどのような目的のもとに妥当と言えるのだろうか．ここではリテラシーレベルの位置付けも踏まえながら，データ分析ツールにはどのような特性が望ましいのかを論じたい．

2　GUI ツールであることの欠点

Excel に限らず GUI 利用のツールでは，データ分析処理においていくつかの問題点を抱えている．

2.1　データ分析の再現性

サイエンスの基本特性の一つに，論じるべき現象の**再現性**がある．データ分析ツールに対してこの再現性を期待するとき，データ分析の最前線でもよく使われている R や Python においては，一連の処理がコマンド列として記述されており，再現性に問題は生じない．学術専門誌の中には，実験データ分析の再現性のために，実データおよび一連の分析手順をいっしょに提出することを義務付ける動きの進んでいる分野もある．

ところが Excel に限らず，GUI 利用のツールではその再現性において問題を生じてしまう．たとえば教育面において，実際に授業の中で実演することを考えてみる．動画あるいはライブの実演を通じて分析処理に関する一連の操作を目にした学生の中に，その操作を正確に再現できない者が相当数いることを多くの指導者が経験しているはずである．一見簡単そうに見えて，次々と進むメニュー選択操作を適切に追いかけることのできていない学生が相当数いるのである．このような特徴は Excel に限った話ではなく，GUI ツール全般に見られることである．

こうした実態を踏まえて，Excel の手引書の多くは，過剰と思えるほどにスクリーンショットを多用している事実もある [2]．

さらに，指導者側と学生のバージョンが異なるときには，画面表示が異なることも珍しくない．

2.2　データの保全性

読み取り専用（Read Only）に設定していないデータが書き換え可能であることは，ツールの利用法が CUI であろうと GUI であろうと生じる．

さらに GUI ツールにおける問題点として，ドラッグ・アンド・ドロップ操作のミスにより書き換えが生じる，あるいは更新が変更したいデータに及ばないという事態が生じやすいことがあげられる．とくに初心者においては，

この操作の安定性はよくない．この現象を称して「コピペ汚染」という言われ方もする．

また Excel では，セルに入力された数値データに対して「日付型」とか「通貨型」などの勝手なデータ型の付与が行われることも珍しくない．

2.3 Excel に関する欠点のまとめ

以上，とくに Excel におけるデータ分析ツールとしての欠点をまとめると次の 3 点がある．

- 分析処理手順の再現性が不十分である
- コピペ汚染を生じる
- データ型の勝手な付与が生じることがある

3 目標をどこに置くか

前節で述べたような欠点があっても，多くの人にとってとっつきやすいという明らかな利点が Excel にはある．だからリテラシーレベルであれば，前記の欠点があっても教育上の利点が上回れば欠点は気にするものではない，と考えてよいだろう．

3.1 分析操作に慣れるのは易しくない GUI ツール

GUI ツール利用での入門期教育において，何を演習の目的にするかによって操作の難易度はもちろん違ってくる．分散の計算を例とすれば，次の 3 つの取り組み方を代表的なものと考えてよいだろう．

1. 分散の計算式をフォローしてみる
2. 分散関数を利用する
3. 分散の計算を含むパッケージ（たとえば「分析ツール」）を使う

これらのいずれの方法を選択する場合でも，事前に表形式のデータを入力済みの Excel シートを使わずに，学生に表形式のデータを作成させる場合には，このデータ作成の作業にかかる時間が無視できないために，授業時間内に終える演習構成を難しくしてしまうことを筆者は経験している [3].

この表形式のデータを扱うという一見単純な作業に困難を抱える学生が相当数いるという実態を踏まえると，ベクトルで数値データを扱うという考え方をもとにする R や Python では，該当する数学分野の苦手感が学習の障壁になるという議論において，ベクトルが表形式データに置き換わっただけの議論にも見える．

3.2 学習支援に時間がかかる GUI ツール

PC を教室利用してデータ分析の演習を行う授業を想定する場合に，何かの課題がうまく行っていない学生を支援するためには，指導者側が学生の PC 画面を閲覧できる状況を作らないと問題解決につなげるのが難しい．

Excel などの GUI ツールでは最初にとりかかるための障壁が低いのは事実であるが,「最初のとりかかりやすさ」を第一の利点として考えるとき，この利点がないと学習を始めることが困難な学生にとっては，データ分析が一通りできるまでの道のりは実は長い時間がかかると考えてよいと思える．

3.3 学習後も処理操作に時間がかかる GUI ツール

仮にデータ分析の処理操作が身についた学生を想定しても，Excel でデータ分析処理をすることは，おおいに操作時間がかかる．R や Python のような CUI ツールではデータ行を書き換えさえすれば一瞬にして分析処理が終わるのに比べるとなおさらである．これは GUI ツールの特性であり仕方のないことである．VBA などによるマクロ化をすれば簡単になるという考え方はもちろんあるだろう．しかし，多くは前述の表形式データを汎用性のあるものにマクロを作れるプログラミングスキルのない者にとって，再利用性の高いマクロを作るのは困難である．

マクロ化に関するこの状況は，ビジネス場面において Excel を使った業務が属人性を抜け出せない問題を生じやすいことにも現れている．RPA のように皮相的で，本質的には業務のデジタル化になっていない方法による業務の合理化が，多くのビジネス場面で利用されているのである．

4 CUI ツールは本当に難しいのか

前節で述べてきた GUI ツール利用に伴う欠点を生じないで授業での実演などを進める方法が一つある．学生側が各自でデータ分析できることを目標とせずに，習熟者による実演デモとして授業利用するのである．うまく進む実演デモを見るだけなら，演習に多くの時間を必要とする学習者にとっても，分析処理は簡単に見えるからである．

この方法であれば，R なら R マークダウンファイル，Python なら Jupyter Notebook ファイルなどで教材提供

することにより，クラウド上にあるこれらのツール（プログラミング言語）の実行環境で簡単に実演できる．しかも，冒頭で述べた「応用基礎レベル」に進む学生にとっても，これらは十分なデータ分析ツールである．

5　結論と将来

以上で述べてきたことから，データ分析の初歩を学習するのに短い時間で済むし，その後も利用価値が高いのは CUI ツールだと考えるがどうだろうか．

またデータ分析初心者にとって，R と Python とでどちらがとっつき易いかと考えれば，プログラムとしてのルールを常に要請する Python よりは，わずか一行で記述できる例題を豊富に提供できる R が易しい．Python を指向するのであれば，応用基礎レベルでプログラミングの基礎を学習したことが役に立つ道筋を用意したい．

将来的にはデータ分析利用者との対話環境が，今後の人工知能技術の進展により，もっと親しみやすいものになるはずである．そのときに，2.3 節に述べたような，処理すべきデータ自体に欠陥を生じやすいツールを使うのか，また GUI と CUI のツールとでどちらのほうが人工知能技術にとって入力情報を把握しやすいか，そのような問題も合わせて考えるとき，CUI ツールの利点はさらに大きくなるだろう．

参考文献

[1] 文部科学省サイトのページ https://www.maff.go.jp/j/kanbo/tizai/brand/attach/pdf/ai-15.pdf，2019

[2] たとえば「できる Excel」シリーズ，インプレス

[3] 鈴木治郎，「Excel で実験する統計学 第 2 版」，ピアソン・エデュケーション，1998