R 言語を用いた再生可能な教学 IR 情報の分析と可視化

西山慶太（専修大学）

１.背景

Institutional Research（以下，IR）は高等教育機関等で行われる様々な活動について，機関内に存在するデータを収集・分析し，その結果に基づいた報告を行うことで，高等教育機関等の改革や改善に寄与することを目的としている．したがって，何らかのデータ分析を行うことは，IR を担当する者にとって必須の事項であり，どのようにしてデータ分析を行うかということが，日々の課題となる．また，IR は非常に学際的な分野であり，データ分析に関する手続きや方法論が確立されているとは言えない．そのため，学術的な作法に習熟している研究者から，学術的なトレーニングを全く受けていない一般の職員まで，幅広い層の人たちが IR に従事しており，個々のスキルや環境に合った方法で業務に従事している．ここで問題になるのが，共通の方法論を持たない IR において，研究結果が妥当なものであり再現可能なものなのかということである．

再現性（Replicability）とは，ある現象が他の研究者が行った研究でも再現されることであり，新規に収集したデータを適用しても同一の結果が得られるということである[1]. IR に当てはめれば，ある機関で確認された現象が，別の機関でも確認できるか否かということになるが，それぞれの機関で異なる機能や形態を持つ高等教育機関においては，一般的な意味での再現性というものを担保することは難しいのではないかと思われる．一方，本稿のタイトルである再生性（Reproducibility）とは研究者自身が前と同じデータを用いて，研究結果等が再生出来ることであり，新規にデータの収集を行わず，前回使用したデータ・セットから同一の結果が再生出来るか否かを指す．これを聞いて至極当然のように再生出来るものであろうと思われるかもしれないが，有名科学雑誌 Nature が 1576 名の研究者に対して行った調査によると，約 70%が他の研究者が結果を再現できず（再現性の問題），50%は自分の研究結果すら再生出来なかった（再生性の問題）という衝撃の事実も明らかになっている[2]. これらの原因としては *p*-hacking や実験計画の問題が挙げられているが，生データから最終的な図までのすべてのステップをリトレース出来るように記録しておくということも重要な要素であるとしている．

このような状況の中，様々なバックグラウンドを持つ人たちによって運営される IR において，再生性は確保されているであろうか．特にデータ分析の工程においては，多種多様なツールが存在し，統計学に関する基礎的な知識がなくとも，統計解析や可視化を行うことが出来るようになっている状況となっているため，アドホックなデータ分析が行われているのではないだろうか[3]. 本稿ではこのような問題意識を前提として，統計解析言語である R 言語とその統合開発環境である RStudio および世界中で開発されている R パッケージを用いて，データの読み込みから，分析，可視化，レポートまでの工程を，一つの解析環境の中で完結する事例を紹介する．

２.R 言語とその周辺環境

R 言語はオープンソース・フリーソフトウェアであり，統計解析に特化したプログラミング言語である[4]．また，R 言語の統合開発環境である RStudio も無償版が RStudio 社から公開されており，こちらも無料で利用することが出来る[5]．R パッケージは R 上で作動する便利な関数の一般名称である．世界中の R ユーザーによって活発に開発が行われており，最新の解析手法も早期に公開されることが多い．R パッケージを利用するには，CRAN（Comprehensive R Archive Network）から R パッケージをダウンロードすればよいだけである．また，CRAN に登録されている R パッケージは一定の審査を経て登録されており，ヘルプ機能から詳細を確認することも出来る．

３.データ分析の工程

まず，一般的なデータ分析のフローを図１に示す．これを見ると，まず，リレーショナル・データベースやテキストデータを，MS Excel などを用いて読み込み，データの加工・整形を行う．その後， MS Excel あるいは各種 BI ツール等を用いて統計解析や可視化を行う．そして，それらの結果をコピー&ペーストし，MS Word を使って書かれたレポートに手動で組み込んでいくという手順である．このデータ分析フローには筆者の主観も入っているが，少なくとも一つのプラットフォームの中で，すべての工程を行うことは稀であると思われる．

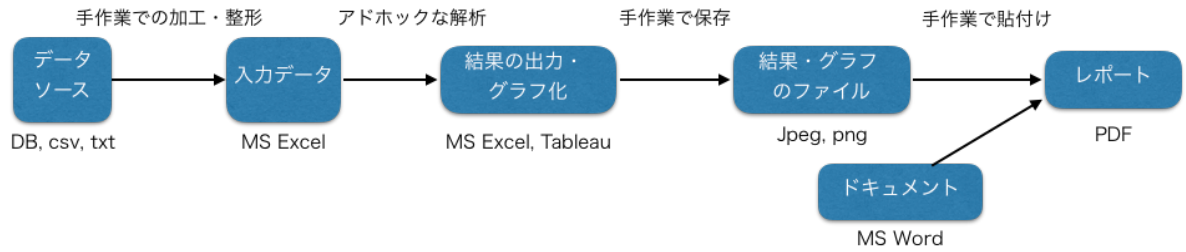

図 １：一般的に行われるデータ分析のフロー（高橋 2018 [3]を基に著者改変）

次に，R 環境を用いたデータ分析フローを図２に示す．ここに示されたデータ分析フローはすべて R 環境の中だけで完結する．具体的には，各種 R パッケージを用いて，データ加工・整形，文字列処理，可視化，統計解析などを行うことになるが，それらすべての工程は R 言語を用いた R スクリプトによって記述され，それらをレポーティングパッケージである R マークダウンパッケージ[6], [7]を用いて統合し，レポートを生成するのである．

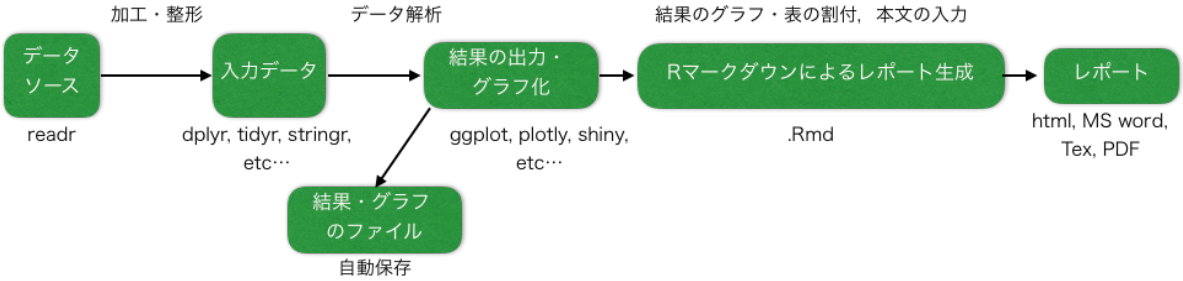

図 ２：R 環境を用いたデータ分析フロー（高橋 2018[3]を基に筆者改変）

４.実際の使用例

１）高等教育の学修支援新制度

本稿では一つの事例として，令和二年度からスタートする高等教育の学修支援新制度[8]の支援対象者の要件（個人要件）等についてデータ分析，可視化およびレポート作成を行う過程を示すこととする．本制度の大学等への進学後の個人要件の内，GPA と修得単位に関する支援打切りおよび警告の条件は表１のとおりである．本稿ではここにある「修得単位数の標準」について，１２４単位を４年で除した３１単位を学年ごとの標準修得単位数として設定する．また，GPA の下位４分の１については，各学部，各学科，学年ごとのグループを一つの集団とし，集団ごとにグループ内順位を算出することとする．

表 1:「高等教育の学修支援新制度」における修得単位数と GPA に係る個人要件

直ちに支援「打ち切り」	「警告」（連続で警告を受けた場合打ち切り）
修得単位数が標準の５割以下の場合	修得単位数が標準の６割以下の場合
−	GPA 等が下位４分の１の場合

２）ダミーデータの生成

この事例で使用するデータ・セットは筆者が作成したダミーデータであり，現実には存在しないデータである．なお，ダミーデータを生成した R スクリプトは参考文献を参照のこと[9]．生成したダミーデータの構造は表１の通りであり，GPA と修得単位数の２つの量的変数と学部，学科，学年の属性を表す変数から構成されている．なお，一つのレコードは学生一人を表現しており，１学年 3000 名を４学年分，合計 12000 名のデータを生成した．

表 ２：ダミーデータの構造(１学年 3,000 名を想定し，12,000 名分のデータ)

変数名	定義	データ範囲	データ型
sID	学籍番号	学部・学科・学年・連番	charactor 型
school	学部名称	７学部	factor 型
department	学科名称	２３学科	factor 型
grade	学年	１～４年次	Integer 型（整数値）
GPA	年度内 GPA	０～４	numeric（小数点第二位まで）
credit	累積修得単位数	０～１６０	Integer 型（整数値）

３）R を用いたデータ分析

本稿では紙面の都合上，R の使用方法に関する細かい事項まで言及することは困難である．したがって，重要だと思われる点に関しては使用方法を記載するが，その他，本文に記載が出来なかった部分に関しては，R マークダウンを使用して作成した分析結果レポートを，参考文献に挙げておくので，そちらを参照してもらいたい[10].

４）学部，学科，学年ごとの GPA 順位の計算

データを R 上に読み込んだ後，学部，学科，学年でグルーピングし，それぞれのグループ内での GPA 順位を計算する．表１のデータ概要を見ると，７学部，２３学科，４学年のデータであるため，９２グループについて同じ処理を行う必要がある．これを MS Excel で行えば，グループごとにシートを分けるなどした後に関数を使うこととなるため，作業量は膨大になり，ミスも発生しやすくなるだろう．しかし，R の dplyr パッケージ[11]を使用すれば図３の R スクリプトのように非常に少ない記述で処理が実行することが出来る．

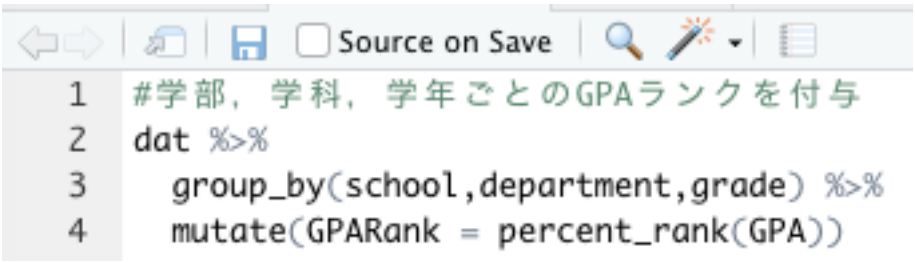

図 3：グループごとの GPA 順位を計算する R スクリプト

このたった四行の R スクリプトは，

１行目　”#”から始まるコメントアウト（メモ）
２行目　csv データを保存した dat というオブジェクト．”%>%”は処理を継続の記法.
３行目　school, department, grade の三つ要素でグルーピングしてください.
４行目　GPARank という列を追加し，グループごとの GPA のランクを計算する.

という処理となる．ここまでが可視化の前段階である前処理に当たる部分である．

５）データの可視化

前処理を行った後，グループごとの GPA と修得単位数の二変数を，ggplot パッケージ[12]を用いて散布図にしたものが次葉の図４である．すべてのグループのグラフを表示すると，その数が膨大になるため，紙面の都合上ここでは１年次のみに絞って掲載している．また，「打ち切り」および「警告」のボーダーラインについては，実線および点線でそれぞれを表現している．なお，この図を作成するために使用している ggplot パッケージや上述の dplyr パッケージは，RStudio 社の開発チームが作成した tidyverse パッケージ[13]というパッケージ群に含まれている．このパッケージ群にはデータの読み込み，前処理，文字列操作，可視化，統計モデリングに係る様々なパッケージが含まれており，R を用いてデータ分析を行う際には必須となる．

６）R マークダウンでのレポート作成

ここまで行った作業はすべて R スクリプトに記録されている．R マークダウンではチャンクという機能を用いて，これらのスクリプトを文書内に表示させることが出来る．また，レポートの本文に当たるテキストについても，R マークダウンファイル内に直接書き込む

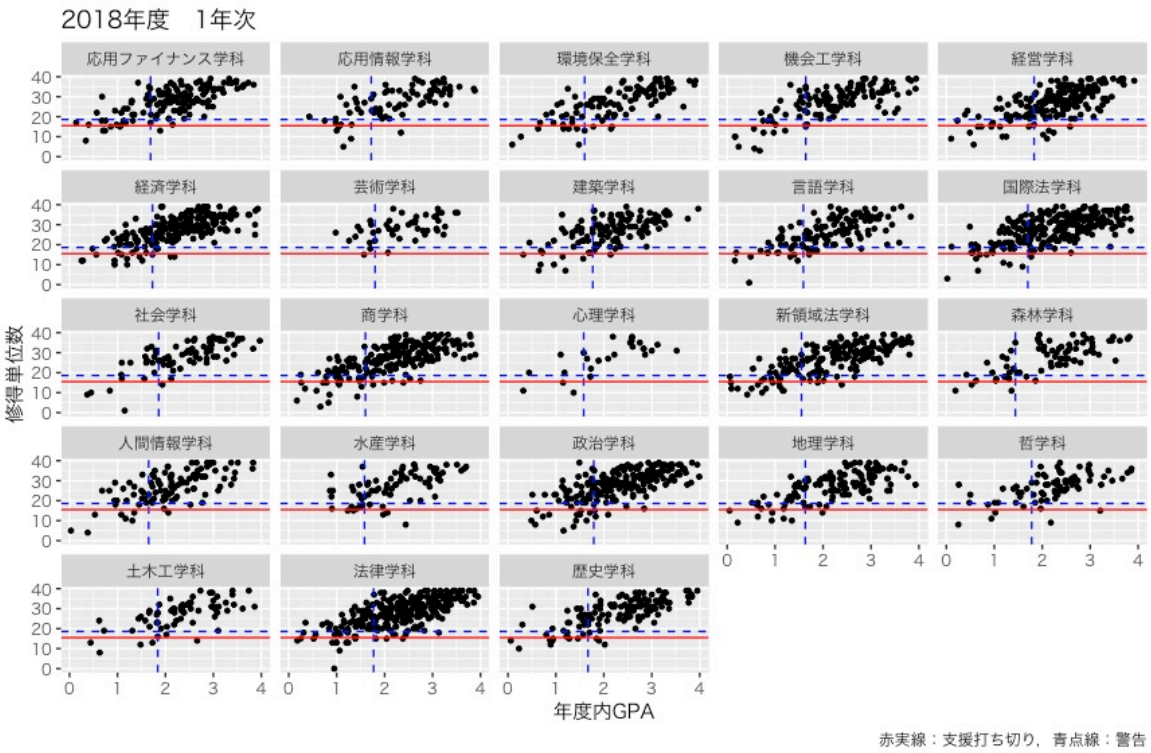

図 4：GPA と修得単位数の散布図および「打ち切り」,「警告」ライン

ことが出来る．これらを組み合わせることで，R 環境で処理したデータ分析の結果と文章を一つのファイルに統合することができる．また, R マークダウンで作成したレポートは，html，PDF, MS word, Tex などの様々な形式で出力することが出来るため，レポートの用途に応じて選択することが可能となる．

５．まとめと考察

ここまで，R 環境を用いて，ダミーデータを生成し，そのデータを加工，可視化，分析したものを，R マークダウンによって統合しレポートをする過程を紹介した．繰り返しになるが，本稿で触れることが出来なかった部分については，すべてのデータ分析工程を統合した R マークダウンから出力された html ファイルを web 上に掲載しておくので，そちらを参照してもらいたい[10]．このように R 環境を利用することで，ほぼすべてのデータ分析フローが一つのソフトウェア上で可能となる．それでは，IR 業務にとってこのことがどのような利点となりうるだろうか．

一つにはルーチン型業務の効率化である．本稿で示した修学支援の要件確認は，今後毎年行うことが必要となるが，R スクリプトや R マークダウンの形式でデータ分析過程を保存しておくことで，次回以降の処理も瞬時に行うことが出来る．これは従来の方法に比べて大きな利点である．もう一つは，基データのフォーマットが統一的に作成できる場合には，学内だけでなく，機関を超えてコードシェアが出来る可能性である．今回の例で取り上げた高等教育の修学支援新制度などは，国内の多くの大学で同じデータ処理を行うことが予想される．したがって，作成したコードをシェアすることができれば，高等教育機関

全体でも業務の効率化が図れる可能性がある.

以上のように R 環境を使用すると様々な利点があるが，実際の運用には問題も残る．一番大きな問題は R 言語がスクリプト言語であるため，修得にある程度のコストを要する点である．IRer にとってデータ分析技術は必須の事項であるが，現状では本稿で紹介した初歩的な分析フローであっても,それを業務レベルで実現できる人材は数少ないと思われる．したがって，再生可能なデータ分析フローを実現し一般的にしていくためには，IRer の人材育成も合わせて行っていく必要があるのではないだろうか.

【参考文献】

[1] 国里愛彦, “統計解析の再現可能性を高める取り組み,” 2017. [Online]. Available: https://www.slideshare.net/YoshihikoKunisato/ss-77835559. [Accessed: 30-Sep-2019].

[2] M. Baker, “1,500 scientists lift the lid on reproducibility,” *Nature*, 2016.

[3] 高橋康介, *再現可能性のすゝめ：RStudioによるデータ解析とレポート作成*. 東京: 共立出版, 2018.

[4] R Core Team, “R: A Language and Environment for Statistical Computing.” Vienna, Austria, 2018.

[5] RStudio Team, “RStudio: Integrated Development Environment for R.” Boston, MA, 2018.

[6] J. J. Allaire, Y. Xie, J. McPherson, J. Luraschi, K. Ushey, A. Atkins, H. Wickham, J. Cheng, W. Chang, and R. Iannone, “rmarkdown: Dynamic Documents for R.” 2018.

[7] Y. Xie, J. J. Allaire, and G. Grolemund, *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman and Hall/CRC, 2018.

[8] “高等教育の修学支援新制度：文部科学省.” [Online]. Available: http://www.mext.go.jp/a_menu/koutou/hutankeigen/index.htm. [Accessed: 01-Oct-2019].

[9] K. Nishiyama, “Data Generating Script,” 2019. [Online]. Available: http://rpubs.com/keita_nishiyama/534598. [Accessed: 02-Oct-2019].

[10] K. Nishiyama, “Rmarkdown Reporting Script,” 2019. [Online]. Available: http://rpubs.com/keita_nishiyama/534617. [Accessed: 02-Oct-2019].

[11] H. Wickham, R. François, L. Henry, and K. Müller, “dplyr: A Grammar of Data Manipulation.” 2019.

[12] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[13] H. Wickham, “tidyverse: Easily Install and Load the ‘Tidyverse.’” 2017.