はじめよう 再現可能なデータ解析

中京大学心理学部 准教授
高橋康介（たかはし　こうすけ）

「再現可能性」というパワーワードがバズって久しいです。心理学の再現可能性は統計の誤用，出版バイアス，プレレジなどが絡み合う複雑な問題です。本稿もまた再現可能性の話ですが，全くややこしい話ではなく，「再現可能なデータ解析をしましょう」という，単純で誰もが今から実践できる話です。誌面の都合でほんのさわりの紹介だけなので，本稿の内容に興味を持った方は，ぜひ拙著『再現可能性のすゝめ』[1]を手にとって見てください。

再現できないデータ解析

実験・調査を行い，データ解析して，論文や報告書，プレゼン資料などをつくります。家に帰るまでが遠足，ではありませんが，解析して終わりではなく誰かに見せるまでがデータ解析です。まず調査・実験データが手元にあります。エクセルなどの表計算ソフトで整形・集計したり，グラフを作成したり，統計ソフトにコピペしたり。統計ソフトではメニューを選んで，オプションをチェックして，実行ボタンをクリック。完成したグラフや統計ソフトの結果をプレゼン資料に貼りつけます。よくみかける光景ですね。表計算ソフトでは画面いっぱいに数値を眺めることができます。見栄えの良いプレゼン資料も完成します。満足や充実感は得られるでしょう。

では再現可能性は得られるでしょうか。プレゼン用の資料が失われたときに，同じものを作り出せるでしょうか。いたるところで込み入った手作業を忠実に再現できない限り，再現可能性は保証されません。そして手作業を忠実に再現することは不可能です。手順を正確にメモしておけばいいとか，「私，失敗しないので」というツワモノもいるかもしれませんが，それは過信です。人間は必ず間違えます。

再現可能なデータ解析の思考モデル

一連の作業はデータからプレゼン資料への「変換」とみなせます。ただし再現できない一回限りの変換です。逆に再現可能な変換とは，同じデータから誰でもいつでも同じプレゼン資料を作り出せるということです。「データ解析が再現可能である」とは，「同じデータを入れれば同じモノに変換してくれる何かが存在する世界」に他なりません。ということは，重要なのは出来上がったプレゼン資料そのもの（変換した結果）ではなくプロセス（変換）の存在です。再現可能なデータ解析とは，プレゼン資料を作ることではなく，再現可能な変換を作ることです。この視点の転換は非常に重要です。

では再現可能な変換には何が必要でしょうか。答えは簡単で，すべてを機械化・自動化するためのレシピです。変換の中からあらゆる手作業を排除し，これと等しい処理をコンピュータで実行可能にすることです。人間は間違えたり逆らったりしますが，機械は命令に忠実です。ですから人がやるべきことは，機械に与える命令（レシピ＝解析プログラムなど）の作成です。結果を得ることではなく命令を整えることこそが再現可能なデータ解析の目的である，という考えを徹底的に頭に叩き込みましょう。料理を作るのが目的なのではなく，誰もが同じ料理を作れるレシピを作成することが目的，ということですね。

例えばNature誌の投稿ガイド[2]には“*We encourage authors to make openly available any code or scripts that would help readers reproduce any data-processing steps.*”と明記されています。ここでのcode or scriptsが「変換のレシピ」に当たります。レシピとデータがあれば，誰でもいつでも同じ結果を再現できるというわけです。

再現可能なデータ解析のメリット

とはいえ，これまで手作業で簡単にやっていたことを機械化する，その心理的障壁は低くありません。ここでは再現可能なデータ解析のメリットを強調しておきます。第一に信頼性の向上です。手作業にミスはつきものですが，機械は間違えません。第二に間違いの検証が挙げられます。プレゼン資料をみて「なんか変」だったとしましょう。手作業の場合は，どこでどのようなミスがあったか検証する術がありません。機械化すればミスも再現されるため，どこが変なのか後から追跡できますし，修正も容易です。第三に作業効率の圧倒的な向上です。手作業の場合，データが増えれば掛け算で作業量が増えます。データ集計にミスが見つかれば，その後の作業は全部やり直し。悪夢のようですが，こんな経験がある人も多いでしょう。機械化されていれば，

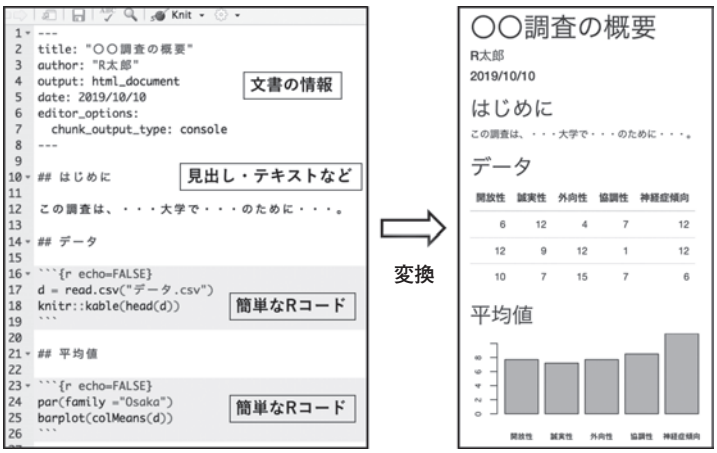
このような悪夢から抜け出すことができるのです。逆に，技術習得のコストを除けば，デメリットはありません。

再現可能なデータ解析を始めよう

どこから始めればいいでしょうか。最近は再現可能なデータ解析のためのツールがたくさんあります。「変換レシピの作成」という視点の転換ができていればどのようなツールを使っても構いませんが，心理学者にとって一番使いやすく汎用性が高い，そして習得のメリットが大きいツールはおそらくR，RStudio，そしてRマークダウンでしょう。

Rは言わずとしれた統計解析用ソフトウェアです。無料です。データの整形から統計解析，図表作成など，簡単なR言語のコードで自由自在にデータ解析を行うことができます。最近では学部教育でRを導入したという話もチラホラと出始めています。RStudioはRを劇的に使いやすくするソフトウェアで，RはツンでRStudioはデレです。特に理由がない限りRStudioとRを使いましょう。

最後にRマークダウンとは何か。一言で表現するのは難しいのですが，プレゼン資料の作成まで再現可能にしてくれるものです。言葉にすると取るに足らないもののようですが，使ってみて初めてわかる，計り知れない恩恵があります。資料作成はデータが増えるほど作業量が増え，混乱の度合いを増します。どのグラフを貼ればいいのか，このグラフに対応する統計結果はどこなのか，この結果を出力したコードはどれなのか……，ファイルを探し回り，あるいはエクセルの中を探し回ってなんとか作り上げた資料。それを，もう一度最初から同じものを作れと言われたら，涙が出てきます。



Rマークダウン　　出力（HTML・ワード・PDFなど）

図1　Rマークダウンの動作イメージ。左がRマークダウンで，テキストや画像とRコードからなるプレゼン資料のタネ。これが自動的に変換されて，コードが集計結果やグラフなどに置き換えられたプレゼン資料が出力される。

Rマークダウンはそんな悩みから人々を救うのです。動作イメージを示しましょう（図1）。左側は手で作成する「Rマークダウン」です。プレゼン資料を生み出すレシピで，文書の情報，注釈テキストなどに加え，Rのコード（コードの実行結果ではなく！）が書かれています。この例では，調査の説明，データ読み込み，表作成，平均値集計とグラフ化の「コード」があります。再現可能な変換のすべてが，この中にあります。あとはRStudioの「knitボタン」というものを押せば右のプレゼン資料が作成されます。何度でも，同じものが作成されます。データが追加されたら？　もう一度ボタンを押すだけです。もとのデータはどこにあるの？　データを読み込むコードを確認するだけです。グラフが変だったら？　グラフを出力するコードを確認するだけです。再び手作業で集計する，結果が入ったファイルを探し回る，エクセルの中を探し回る，もうそんな必要はありません。

このようにRマークダウンとは「同じデータを入れれば同じモノに変換してくれる何かが存在する世界」です。ここでRマークダウンの実力を説明するには全く誌面が足りません。本稿を読んで再現可能なデータ解析に目覚めた方は，ぜひ『再現可能性のすゝめ』を手にとって，その素晴らしい世界に足を踏み入れてみてください。

1　高橋康介（2018）『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』（シリーズWonderful R／石田基広 監修・市川太祐・高橋康介・高柳慎一・福島真太朗・松浦健太郎 編）共立出版　https://www.kyoritsu-pub.co.jp/bookdetail/9784320112438

2　NPG: https://www.nature.com/sdata/publish/submission-guidelines

Profile — 高橋康介
京都大学大学院情報学研究科修了。博士（情報学）。JSPS特別研究員SPD，東京大学先端科学技術研究センター特任助教などを経て現職。専門は認知心理学。