Wonderful R［3］巻 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―
高橋 康介著・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
http://www.kyoritsu-pub.co.jp/bookdetail/9784320112438

序

Rによるデータ解析について解説した書籍は数多くあるが、「データ解析という作業」について解説した書籍は少ない。本書はRStudioを活用して**再現可能なデータ解析とレポート作成**を身につけるための一冊である。

再現可能性とは何だろうか。

科学における再現可能性とは、同じ環境で同じ実験や調査を行えば同じデータが得られて同じ結論が導かれるということである。たとえ一度は華々しく脚光を浴びた研究成果でも、再現性が低ければいつかは淘汰されることになる。

近年、一部の科学分野は再現可能性の危機に直面している。例えば、過去の心理学の研究のなかで結果が再現できるものは40%に満たないという衝撃的な論文が、2015年にScience誌で発表された[1) 2)]。このような背景から、科学者の間では既存の研究の再現可能性を評価する追試研究の重要性が見直されつつある。また、事前登録制度の活用など、なんとかして再現可能性の高い研究を行おうという気運が強まっている。

このような再現可能性の問題は一部の科学分野のものだけではない。データ解析に携わる人なら、自分で解析した結果の再現可能性について考えたことがある人は多いだろう。解析した結果を再現できなくて冷や汗をかいたことがある人も少なくないかもしれない。

第1章で詳しく紹介するが、データ解析は再現可能性が問題となる代表的な事例である。手にしたデータ解析結果を翌日に再現できないなら、誰もそのデータ解析の結果をビジネスの意思決定に活かしたいとは思わないだろうし、そのような再現できない結果を報告してくる人を信用しようとは思わない。

このように、データ解析に携わる人にとって再現可能性を高めることが重要な課題であることに間違いはない。では、どうすれば再現可能なデータ解析とレポート作成を行うことができるか。その問いに答えるのが本書の役目である。

さて、本書と同じく共立出版から「Useful R」第9巻として『ドキュメント・プレゼンテーション生成』を刊行したのが2014年6月である。『ドキュメント・プレゼンテーション生成』では、Rによる再現可能なレポート作成について解説したが、当時はコンソール上で**knitr**パッケージを直接操作する方法がメインで、初心者には多少ハードルが高かったかもしれない。

---

1) http://science.sciencemag.org/content/349/6251/aac4716

2) 心理学における再現可能性の問題については心理学評論の特集号『心理学の再現可能性：我々はどこから来たのか　我々は何者か　我々はどこへ行くのか』に詳しい。http://team1mile.com/sjpr59-1/ にてすべての論文がオープンアクセスで公開されている。

しかし『ドキュメント・プレゼンテーション生成』から3年余りを経て、Rを取り巻く環境は大いに進歩している。その立役者が、何といっても近年台頭してきたRStudioである。

実は再現可能性という観点からは、『ドキュメント・プレゼンテーション生成』の当時と比べて現在でも**やれること**はそう多くは変わっていない。変わったのは**やりやすさ**である。

RStudioにより、再現可能性の高いデータ解析とレポート作成のための作業プロセスを誰でも容易に導入できる。RStudioにより、これらの作業の効率を大幅に上げることができる。今では再現可能なデータ解析とレポート作成は、一部のエキスパートだけでなく誰もが使える技術となっている。このような理由から、『ドキュメント・プレゼンテーション生成』の続編という位置づけで本書を刊行するに至った。

Rの初心者からエキスパートまで、まだ再現可能な作業形態を導入してない場合には、本書によって再現可能なデータ解析とレポート作成の意義を感じ取って、実際に自分の作業の中に導入してほしい。すでに再現可能な作業形態を導入している場合にも、本書によってさらにスキルアップして、再現可能性を高めて、作業効率を上げてほしい。すべての読者が、RStudioによる再現可能なデータ解析とレポート作成の恩恵を享受できるようになれば、筆者としてこの上ない幸せである。

謝辞

本書の刊行にあたり、脱稿を気長に待って頂いた共立出版の石井徹也さん、大谷早紀さん、監修として厳しい締切を設けるとともに暖かく執筆を見守って頂いた石田基広先生、レビューをして頂いた編集委員の皆様、RStudioやRマークダウンに関して大変参考になる資料を公開してくださっている@kazutanさん こと前田和寛氏(比治山大学短期大学部)、そしてなんとなくr-wakalangの皆様に厚くお礼申し上げたい。

本書の構成と対象とする読者

本書の構成は次のとおりである。

- 第1章: データ解析とレポート作成における「再現可能性」の紹介や意義などの解説。
- 第2章: RStudioの紹介と基本的な機能、操作の解説。
- 第3章: RStudioによる再現可能なデータ解析の解説。
- 第4章: Rマークダウンによる再現可能なレポート作成の解説。
- 第5章: Rマークダウンの多彩な表現力を活かす手法の解説。
- 第6章: さらなる再現可能性の向上に向けた手法の解説。
- 第7章: RStudioを使いこなすための機能の解説。
- 付録A: マークダウン記法。

- 付録 B: R マークダウンのオプション (チャンクオプションとパッケージオプション)。

第 1 章では初心者を対象に「再現可能性」について解説しているが、中上級のユーザも確認の意味で、ぜひ読んでみてほしい。

第 2 章は RStudio の紹介、基本機能、操作方法などの説明である。すでに RStudio を使いこなしている場合には、第 2 章は読み飛ばしても構わない。

第 3 章と第 4 章が本書の中心である。第 1 章で紹介するように、再現可能性はデータ解析とレポート作成の 2 段階に分けることができる。第 3 章ではデータ解析フローで再現可能性を高める方法を解説する。スクリプトの利用に始まり、データ読み込みの自動化や解析結果の保存の自動化などの解説、また、プロジェクトなど、再現可能性を高めるための RStudio の機能も紹介する。現在、R スクリプトを使わずにアドホック[3] なデータ解析を行っている場合は、まずは R スクリプトで再現可能なデータ解析を行うことを身につけてほしい。

第 4 章ではレポート作成フローに R マークダウンを導入することで再現可能性を高める方法を解説する。R マークダウンファイルの書き方やレポートの生成方法などを紹介する。第 3 章と第 4 章の内容を身につければ、再現可能なデータ解析とレポート作成を実践できるようになる。

第 5 章から第 7 章までは応用的な内容である。第 5 章では、R マークダウンによるプレゼンや本の作成、**htmlwidgets** ベースの可視化など、R マークダウンの多彩な表現力を活かす術を紹介する。一歩進んだレポート作成技術を身につけたい場合には目を通すとよいだろう。第 6 章では、バージョン管理システムやパッケージ環境の再現など、データ解析やレポート作成の再現可能性をさらに向上させる術を紹介する。第 7 章では、デバッグやコード診断など、RStudio のディープな世界を紹介する。

付録 A にはマークダウン記法を、付録 B には R マークダウンのオプションを、リファレンスとして使えるように一覧できる形で掲載しているので、必要に応じて参考にしてほしい。

本書で想定する読者層について説明しよう。

本書では R そのものの使い方や文法、関数などの解説はしないので、少なくとも R でデータ解析をこなすことができる、またはその経験があることが必須である。R は使えるがアドホックな解析しかしていないので限界を感じているという人から、R マークダウンを使って再現可能なデータ解析とレポート作成を行っているがクオリティをさらに高めたいという人まで、幅広い読者層に有益な情報を掲載したつもりである。

各スキルレベルに対応する章は次のとおりである。自分のスキルに合わせて読み進めてほしい。

- R は使えるが、そろそろ RStudio を使い始めたい→第 2 章。
- アドホックな解析に限界を感じているので、スクリプトを使った再現可能なデータ解析を学びたい→第 3 章。
- レポートの作成でワープロソフトやスライド作成ソフトへのコピペに疲弊しているので、レポートの作成に R マークダウンを使ってみたい→第 4 章。
- すでに R マークダウンを使っているが、さらにかっこいいレポートを作成したい→第 5 章。
- すでに再現可能性を意識してデータ解析とレポート作成を行っているが、さらに再現可能性を高めたい→第 6 章。

[3]「アドホック」の意味については第 1 章で解説している。

- RStudio を十分に活用していない気がするので、RStudio ウィザードになって仕事の効率を上げたい→第 7 章。

本書執筆時の環境

本書ではコードやスクリプトなどは黒枠、R の出力 (実行結果) はグレー領域で示す。

```
コード・スクリプト・R マークダウンなど
```

実行結果

本書執筆時 (2017 年 4 月) の実行環境は次のとおりである。

```
# RStudio のバージョン情報
rstudioapi::versionInfo()$version
```

```
## [1] '1.0.143'
```

```
# R の情報
sessionInfo()
```

```
## R version 3.4.0 (2017-04-21)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Sierra 10.12.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib
##
## locale:
## [1] ja_JP.UTF-8/ja_JP.UTF-8/ja_JP.UTF-8/C/ja_JP.UTF-8/ja_JP.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] magrittr_1.5 knitr_1.17
##
## loaded via a namespace (and not attached):
##  [1] compiler_3.4.0  backports_1.1.1 bookdown_0.5    rprojroot_1.2
##  [5] tools_3.4.0     htmltools_0.3.6 rstudioapi_0.7  yaml_2.1.14
##  [9] Rcpp_0.12.13    stringi_1.1.5   rmarkdown_1.7   stringr_1.2.0
```

```
## [13] digest_0.6.12   evaluate_0.10.1
```

本書の内容は、RStudio バージョン 1.0 に基づいている。しかし、筆者の本書の執筆開始から脱稿まで半年余りが経過してしまい、その間に RStudio バージョン 1.1 がリリースされてしまった。本書で触れる内容に関して、大きな変化はないが、必要に応じてバージョン 1.1 の変更点に関する注を記した。

また、R マークダウンで利用されているドキュメント変換ツール Pandoc のバージョンが 1 から 2 へと上がっている。本書の内容で修正が必要な点についてはサポートサイトにて情報提供する予定である。

RStudio 開発陣からのメッセージ

本書を刊行するにあたり、RStudio Inc. のメンバーである Hadley Wickham 氏 (**tidyverse** パッケージの提唱者) と Yihui Xie 氏 (**knitr** パッケージの開発者) から、日本の読者に向けてのメッセージを寄せてもらったので、ここに抄訳を載せておく。筆者のことを少しばかり褒め過ぎな感があるが、ご愛嬌ということで容赦願いたい。

まずは Hadley からのメッセージ。

> 再現可能性は現在のデータ解析のなかで極めて重要なスキルであり、良い研究を行うための土台となるものです。この本を手に取った皆さんは、何も心配しなくていいでしょう。kohske (注: 筆者のこと) は ggplot2 の開発にも貢献している優れた R プログラマーです。kohske と一緒に再現可能性を習得しましょう。頑張ってください。
>
> Hadley
>
> PS. rpubs.com には kohske が作った私のお気に入りの「再現可能な R スクリプト」が 2 つあります: `https://rpubs.com/kohske/64032` と `https://rpubs.com/kohske/211993` です;)

続いて、Yihui Xie からのメッセージ。

> kohske が RStudio と R マークダウンについての本を出すということで、とても嬉しく思っています。私は日本語はわかりませんが、日本文化にとても興味があります。和風のインテリアデザインや一期一会のような禅の哲学など、日本文化の精神がとても好きです。バドミントンをやるので、日本選手の試合を見るのも好きです。そしてとりわけ、日本のアニメ、Naruto のとんでもない想像力が大好きです。中国語のウェブサイト[4] で Naruto についてよく語っています。私が作った R パッケージの一つ **xaringan** は、Naruto の寫輪眼から名付けたものです。私が作った R パッケージについて日本のユーザがツイッターで議論しているのを、翻訳ツールでたまに見ています。私の人生の目標の一つは、いつか Masashi Kishimoto(訳注: Naruto の作者) に R マークダウンを使ってもらうこ

[4] `https://Yihui.name/cn/`

とです。旅行は好きではありませんが、日本に行きたいと思っています。

最初に kohske と関わったのは、**formatR** パッケージの開発です。kohske は `tidy_source` 関数に、=を自動的に<-に変換するという重要な機能を加えました。私はとても驚き、日本のプログラマーの技術の高さを実感しました。印象に残っている日本のプログラマーがあと 2 人います。@yutannihilation 氏は **knitr** パッケージの開発に貢献してくれています。@kazutan 氏はいろいろな情報をツイッターに上げてくれます。とても感謝しています。

再現可能性は重要です。このことは、この本を通じて kohske が十分に説明してくれると確信しています。**knitr** と R マークダウンはコードと文書をまとめ上げるために開発されました。これによって、レポート作成での統計解析の信頼性と簡便性が高められます。RStudio は R の開発者やユーザに効率的な作業環境を提供するようにデザインされています。私は昔は Emacs ユーザでしたが、後になって RStudio と R の組み合わせがもたらす深みに気づきました。よく考えられた機能がたくさんあり、これは日本文化の精神にもよく合っていると感じます。読者の皆さんも同じように感じてくれることを願っています。そして RStudio と R マークダウンを楽しんでくれることを心から願っています!

RStudio チームの活動

RStudio という名称は、通常は本書で紹介する R 用 IDE というアプリケーションを指すことが多いだろう。しかし RStudio の開発を行っている RStudio Inc. がさまざまなサービスや製品開発を手がけるようになってからは、RStudio という名称が RStudio 開発チームのことを指すようにもなってきた。それほどまでに、RStudio チームは R 界隈において存在感を増している。本書の内容に入る前に、RStudio チームという組織、そして RStudio チームが開発するサービスや製品を紹介しておこう。

RStudio Inc. は JJ Allaire 氏 (ColdFusion の開発者) により 2008 年に設立され、以降、RStudio の開発を中心に、R に関連する製品開発やサービスの提供を進めている。RStudio チームの規模は日に日に大きくなり、現在では 40 名以上のエンジニアや研究者が所属している。2012 年夏には **ggplot2** パッケージの開発者としても有名で、当時から日本でも人気が高かった Hadley Wickham 氏が RStudio チームに加わり[5]、巷を賑わせたことは記憶に新しい。

RStudio 以外の代表的な製品やサービスには次のようなものがある。

- Shiny[6]: データ解析と可視化のためウェブアプリケーションを作成するパッケージ。本書では解析結果のアウトプットとして旧来のレポートやプレゼンを想定しているが、アウトプットとして Shiny によるウェブアプリケーションを作成する場面も増えてきており、今後はますます発展していくだろう。ローカル環境でも動かすことができるが、RStudio では Shiny アプリケーションをサーバ上で動かすための Shiny Server (無償) や Shiny Server Pro (有償) といっ

[5] https://blog.rstudio.org/2012/08/20/welcome-hadley-winston-and-garrett/
[6] https://www.rstudio.com/products/shiny/

た製品[7]、Shinyapps.io[8] という無償で手軽に使えるウェブアプリケーション動作環境を提供している。

- RStudio Connect[9] : RStudio チームが提供するデータ解析に関するさまざまなサービス (Shiny ウェブアプリケーション、R マークダウンレポート、ダッシュボード、グラフなどなど) をチームで共有するための有償サービス。RStudio の **[Publish]** ボタンから簡単に使える。
- RPubs[10] : R マークダウンで作成したレポートを公開できる無償スペース。すべて公開が前提なので業務での利用は避けた方がよいだろう。

また R を快適に、便利に、ストレスなく使うためのさまざまなパッケージを開発、公開している[11]。一部のパッケージ群は **tidyverse** という概念[12] の下に集約されつつある。RStudio が開発に関わっている **tidyverse** のパッケージを以下に紹介しておこう。

- **ggplot2**: 言わずと知れた可視化パッケージ。
- **dplyr**: データ操作。
- **tidyr**: 直感的で効率的なデータ整形。
- **readr**: 表形式データの読み書きのためのパッケージ。
- **purrr**: 関数型プログラミングのサポート。
- **tibble**: データフレームの拡張。
- **hms**: 時間 (時分秒) の処理。
- **stringr**: 文字列処理。
- **lubridate**: 日付と時間の処理。
- **forcats**: 因子 (`factor`) 型データの処理。
- **haven**: R 以外の形式のデータファイルの読み書き。
- **readxl**: Excel 形式のデータの読み込み。

製品、サービス、パッケージの開発以外にも、RStudio::conf というカンファレンス[13] や Webinar と呼ばれるウェブ上で行うセミナーの開催[14]、製品のリリース情報や RStudio チームに加わったメンバーの情報などが掲載される開発者ブログ[15] や R Views[16] という RStudio 製品の導入事例を含む雑多な楽しい記事を掲載するブログを公開している。

RStudio チームの活動にはこれからも目が離せない。最新の情報をフォローしたい場合は、以下のサイトやツイッターをチェックしよう。

- RStudio の公式ツイッター (@rstudio): `https://twitter.com/rstudio`
- RStudio の中の人がつぶやく tips (@rstudiotips): `https://twitter.com/rstudiotips`
- RStudio の開発者ブログ: `https://blog.rstudio.org/`

---

[7] `https://www.rstudio.com/products/shiny-server-pro/`
[8] `https://www.rstudio.com/products/shinyapps/`
[9] `https://www.rstudio.com/products/connect/`
[10] `http://rpubs.com/`
[11] `https://www.rstudio.com/products/rpackages/`
[12] `http://tidyverse.org/`
[13] `https://www.rstudio.com/conference/` 。RStudio チームによるワークショップなどもある。
[14] `https://www.rstudio.com/resources/webinars/`
[15] `https://blog.rstudio.org/`
[16] `https://rviews.rstudio.com/`

- RStudio の情報をいち早く捕捉してアナウンスするアカウント (@kazutan 、非 bot): `https://twitter.com/kazutan`

参考になる情報源

本書を読み進めながら RStudio による再現可能なデータ解析とレポート作成を実践する上で、参考になる情報源を紹介しておこう。とくにコミュニティサイトは心強い味方である。一人でやろうとすると心が折れてしまうことでも、仲間がいれば乗り越えられる。有効に活用して、再現可能なデータ解析とレポート作成の技術を身につけるとともに、一緒に R コミュニティを盛り上げてほしい。

- RStudio のチートシート[17] 。公式は英語だが、下の方にスクロールしていくと RStuido IDE、R Markdown などについては日本語翻訳もあるので、印刷して手元に置いておこう。
- r-wakalang[18] 。slack 上の日本語の R コミュニティ (解説[19] と登録サイト[20])。どんな質問でも気軽にできる。むしろ回答したい人が質問に飢えている状態と表現する方が正しい。2017 年 5 月現在のユーザ数は 400 人弱。
- stackoverflow の R 関連タグ[21] 。言わずと知れた QA サイト。大抵のことは解決する (ただし英語)。ちなみに筆者も昔はよく出没していたので、R タグの金バッジを持っている (プチ自慢)[22]。

商標について

17) `https://www.rstudio.com/resources/cheatsheets/`
18) `https://r-wakalang.slack.com/`
19) `http://qiita.com/uri/items/5583e91bb5301ed5a4ba`
20) `https://r-wakalang.herokuapp.com/`
21) `http://stackoverflow.com/questions/tagged/r`
22) `https://stackoverflow.com/users/314020/kohske`