（ひとりで学べる）実践Ｒケモ・マテリアル・データサイエンス
～付録Ｒスクリプト付き～
Practical Chemo&Material Data Science with R Script

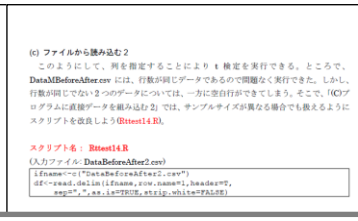
- **奈良先端大　計算システムズ生物学研究室（金谷研究室）メンバーによる執筆！**
- **金谷研究室で作成した、自由に使えるＲスクリプト付き！**
- **機械学習・統計学をいかに習得するかに焦点を当て、化学・マテリアル化学の具体的データを取り入れ、Ｒ言語におけるプログラム例を基に解説！**
- **公開されたＲパッケージを活用するための４つのステップに沿って説明！**
- **一連の流れをひとりでも学べるように編集！研究室に１冊必携の書！**

＜発行要項＞
■発行：2020 年 12 月 30 日
■著者：金谷　重彦 他
■定価：110,000 円(税込)(付録 CD 付)
※CD 中の書籍 PDF は著作権で保護
■体裁：A4 判・並製・308 頁・カラー
■編集・発行：(株)シーエムシー・リサーチ
■ISBN 978-4-904482-95-7

＝　刊行にあたって　＝

第１のパラダイムでは「仮説検定」の発想で経験科学が進展した。そして、ニュートンの法則に代表される第２のパラダイムでは、定量性を考慮した理論科学、計算機の発明により第３のパラダイムとして複雑な現象をシミュレーションに再現する科学へ、そして、現在、第４のパラダイム、として豊富なデータを活用したビッグ・データ・サイエンスが生まれた。ここでは、さらに社会実装も考慮されるようになった。一方、2012 年、ハーバード・ビジネス・レビュー誌がデータサイエンスを「21 世紀で最もカッコいい仕事」と位置づけたことから、注目を集めるようになった。では、これを具体的に化学・マテリアル科学の領域で進めるにはどうしたいいだろうか。実践として必要とされることは、ターゲットとする分野の知識、プログラミングとして必要とされるデータ解析技術としては、データの収集、機械学習と解析結果の評価法（統計学）である。いままで、これらは地道に情報科学の各分野が基盤技術を確立してきた。しかし、オープンサイエンスの時代になり、これらのプログラムおよびデータについてもデータベースとして公開されるようになってきた。ではユーザーとしてこれらのプログラムとデータを活用し、新たな知見を獲得し、さらに社会実装することを目指すことになる。社会実装というと大げさに聞こえるが、企業であれば新たな製品を開発するということへつなげることであり、結局のところ、いま世の中に定着し始めた Sustainable Development Goals（SDGs）に向けた取り組みということへ帰着する。

本書では、このような背景を考慮しながらも、まず「機械学習、統計学をいかに実践的に習得するか？」に焦点を当て、化学・マテリアル科学の具体的データを取り入れ、R 言語におけるプログラム例(約 75 スクリプト)をもとに解説した。R 言語にはさまざま解析用途に応じたパッケージが公開されている。そこで、本書では、＜1. データの入力＞、ファイルからのデータを入力、＜2. データの整形＞、目的に応じて入力データから必要な項目の抽出、＜3. データ解析＞、抽出されたデータを目的にあった関数・パッケージ(統計、多変量解析、機械学習などの関数)に入力し、解析結果を得る。＜4. 解析結果の表示・出力＞　解析結果をもとにグラフに表示する、あるいは、ファイルへ出力する。という４つのステップに沿って、Ｒスクリプトを作成することを説明した。特に、データの整形について本書全体を通して説明を加えた。また、さまざまなパッケージの活用法も習得できるように配慮した。エディタ RStudio、R のインストール、プログラミングの基礎、化学構造からの特徴表現(分子記述子)による多変量データ解析(機械学習により回帰モデル、分類モデル)、妥当性・汎化性能評価という一連の流れを、ひとりでも学べるように本テキスト「(ひとりで学べる) 実践Ｒケモ・マテリアル・データサイエンス～付録Ｒスクリプト付き～」を作成した。また、(株) シーエムシー・リサーチでは、講習会も企画しているそうであるので、これもご活用いただくと、さらに理解が深まり実践的活用への自信もつくと思う。　　金谷　重彦


(c) ファイルから読み込む 2

このようにして、列を指定することにより t 検定を実行できる。ところで、DataMBeforeAfter.csv には、行数が同じデータであるので問題なく実行できた。しかし、行数が同じでない 2 つのデータについては、一方に空白ができてしまう。そこで、(C)プログラムに直接データを組み込む 2) では、サンプルサイズが異なる場合でも使えるようにスクリプトを改良しよう(Rttest14.R)。

スクリプト名：Rttest14.R

(入力ファイル: DataBeforeAfter2.csv)

```
ifname<-c("DataBeforeAfter2.csv")
df<-read.delim(ifname,row.names=1,header=T,
    sep=",",as.is=TRUE,strip.white=FALSE)
```

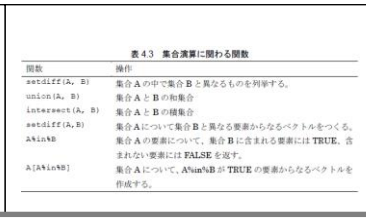


表 4.3　集合演算に関わる関数

関数	操作
setdiff(A, B)	集合 A の中で集合 B と異なるものを列挙する。
union(A, B)	集合 A と B の和集合
intersect(A, B)	集合 A と B の積集合
setdiff(A,B)	集合 A について集合 B と異なる要素からなるベクトルをつくる。
A%in%B	集合 A の要素について、集合 B に含まれる要素には TRUE、含まれない要素には FALSE を返す。
A[A%in%B]	集合 A について、A%in%B が TRUE の要素からなるベクトルを作成する。

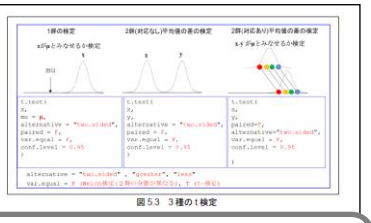


図 5.3　3 種の t 検定

付録 CD ご利用にあたって：付録 CD には、書籍内容の PDF と金谷研で開発したＲスクリプトのフォルダが入っています。Ｒスクリプトは著作権フリーですが、書籍の PDF は著作権で保護されておりますので、ご購入いただきました研究室や部署・部門の範囲でのご利用に限定させていただきます。無断転載・シェアは固くお断りいたします。

注文書		メルマガ会員の登録	登録済み ／ 登録希望
品　名	(ひとりで学べる)実践 R ケモ・マテリアル・データサイエンス	価　格	書籍(付録 CD 付)：100,000 円(税込 110,000 円) ※メルマガ会員は定価の 10%OFF
会社名		ＴＥＬ	
部課名		ＦＡＸ	
お名前		E-mail	
住　所	〒		

お申込み・お問合せ

編集発行：
㈱シーエムシー・リサーチ
101-0054
東京都千代田区神田錦町
２－７　東和錦町ビル３F

ＴＥＬ：０３（３２９３）７０５３
ＦＡＸ：０３（３２９１）５７８９
URL:https://cmcre.com
E-mail：re@cmcre.com

＊書籍はご注文を受けた翌営業日に納品書・請求書とともに送付します。＊お支払いは請求書指定口座に納品日の翌月末日までに振り込みでお願いします。

構成および内容

著者一覧

黄 銘　　奈良先端科学技術大学院大学　先端科学技術研究科　情報科学領域　計算システムズ生物学研究室　助教
小野　直亮　奈良先端科学技術大学院大学　先端科学技術研究科　情報科学領域　計算システムズ生物学研究室　准教授
　　奈良先端科学技術大学院大学　データサイエンス創造センター　准教授
モハマド　アルタフル　アミン
　　奈良先端科学技術大学院大学　先端科学技術研究科　情報科学領域　計算システムズ生物学研究室　准教授
金谷 重彦　　奈良先端科学技術大学院大学　先端科学技術研究科　情報科学領域　計算システムズ生物学研究室　教授