URA, 知の創出センター連携企画 クワトロセミナー（第12 回）
2016 年5 月11 日（水）16:30 - 18:00
東北大学川内南キャンパス 文科系総合研究棟11 階 大会議室

統計解析環境Rを用いたデータ解析

Statistical Computing and Data Analysis with R

医学系研究科循環器EBM 開発学寄付口座

宮田　敏
miyata@cardio.med.tohoku.ac.jp

Agenda:

1. 統計解析環境Rとは

2. Rを使うその前に　ーデータ解析はじめの一歩ー

3. Rを使ったデータ解析の実際

1. 統計解析環境Rとは

統計解析環境Rとは，統計計算とグラフィックスのための言語・環境である．

- インターネット上で配布されるオープンソースのフリーウエア．プログラムの複製，改良，再頒布が可能．
- 多様な統計手法とグラフィックスを提供．柔軟な拡張が可能．
- 高い信頼性．（FDAへの申請にも利用可能）
- 日々拡張される新機能と，パッケージとして提供される最先端の統計手法．

1. 統計解析環境Rとは

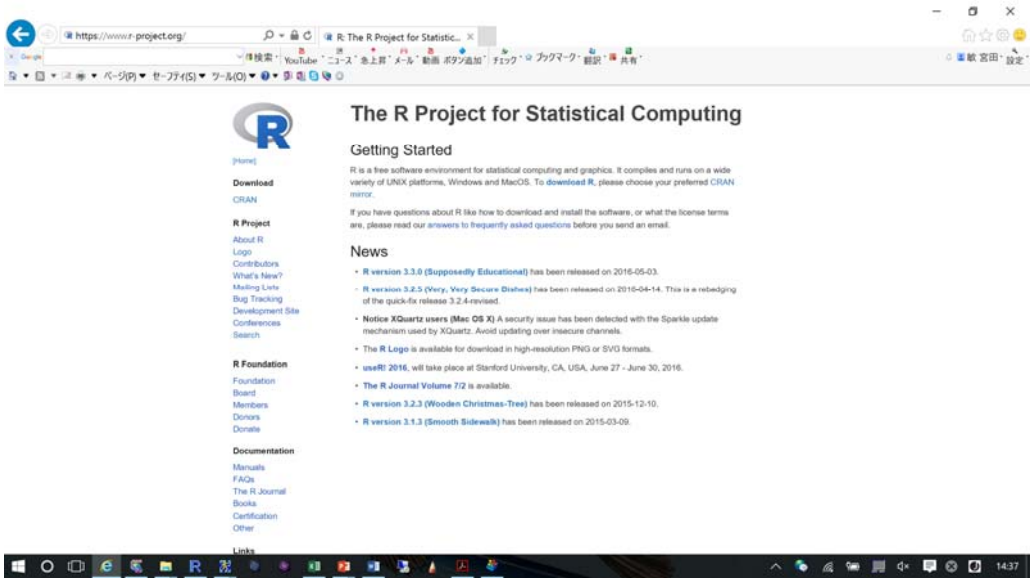
The R project for Statistical Computing: Rの入手先
https://www.r-project.org/

Reference:

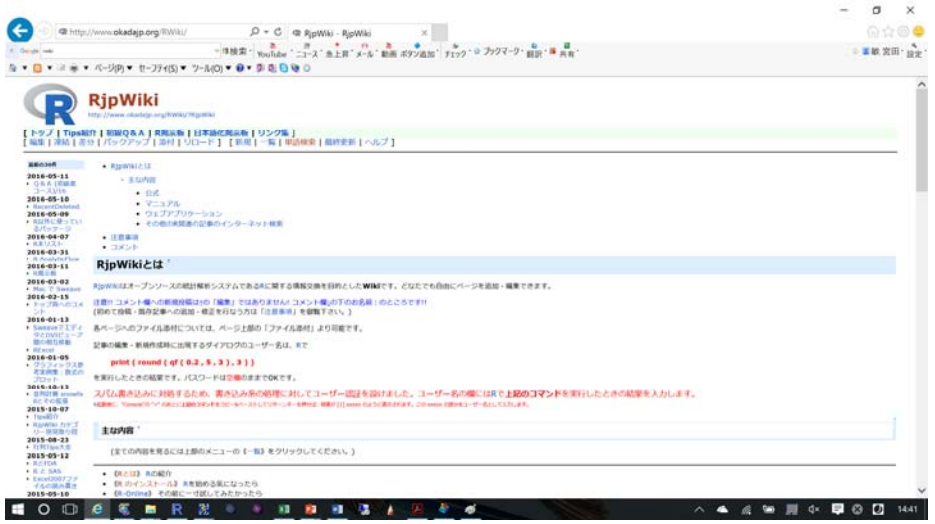
- R Tips: 舟尾 暢男「The R Tipsーデータ解析環境Rの基本技・グラフィックス活用集」
  http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html

- RjpWiki: http://www.okadajp.org/RWiki/

Rは対話型環境からコマンドを入力して利用する．ある程度のプログラミングが必要．


The R Project for Statistical Computing
Getting Started
News


R-Tips


RjpWiki
RjpWikiとは

1. 統計解析環境Rとは

Rはプログラミングを必要とすることが、普及の妨げとなった．マウスを使ってRを使用できるGUI (graphical user interface) が提供されている．

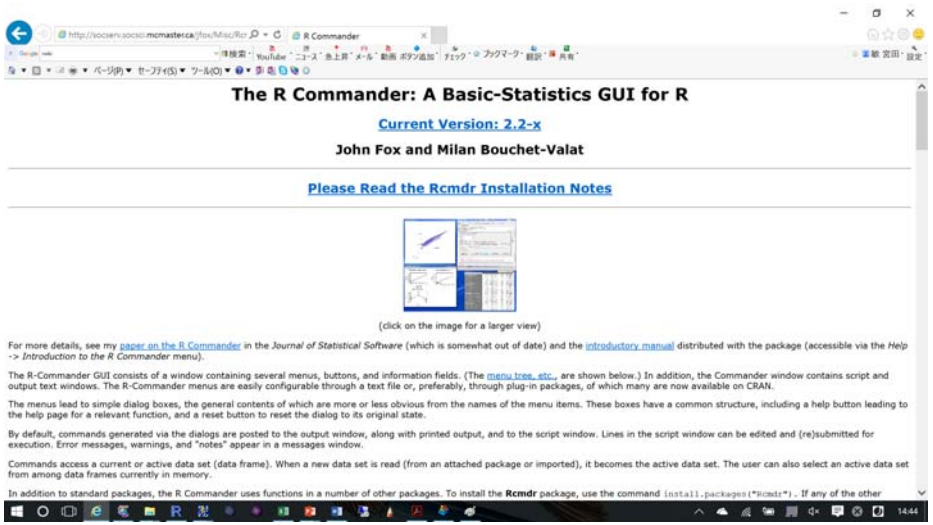
R commander: RのGUIラッパー

http://socserv.socsci.mcmaster.ca/jfox/Misc/Rcmdr/

http://www.ec.kansai-u.ac.jp/user/arakit/R.html

EZR (Easy R): 医学系に強い

http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmed.html


The R Commander: A Basic-Statistics GUI for R
Current Version: 2.2-x
John Fox and Milan Bouchet-Valat
Please Read the Rcmdr Installation Notes
(click on the image for a larger view)
For more details, see my paper on the R Commander in the Journal of Statistical Software (which is somewhat out of date) and the introductory manual distributed with the package (accessible via the Help -> Introduction to the R Commander menu).
The R-Commander GUI consists of a window containing several menus, buttons, and information fields. (The menu tree, etc., are shown below.) In addition, the Commander window contains script and output text windows. The R-Commander menus are easily configurable through a text file or, preferably, through plug-in packages, of which many are now available on CRAN.
The menus lead to simple dialog boxes, the general contents of which are more or less obvious from the names of the menu items. These boxes have a common structure, including a help button leading to the help page for a relevant function, and a reset button to reset the dialog to its original state.
By default, commands generated via the dialogs are posted to the output window, along with printed output, and to the script window. Lines in the script window can be edited and (re)submitted for execution. Error messages, warnings, and "notes" appear in a messages window.
Commands access a current or active data set (data frame). When a new data set is read (from an attached package or imported), it becomes the active data set. The user can also select an active data set from among data frames currently in memory.
In addition to standard packages, the R Commander uses functions in a number of other packages. To install the Rcmdr package, use the command install.packages("Rcmdr"). If any of the other


自治医科大学附属さいたま医療センター
血液科のホームページへようこそ。
自治医科大学附属さいたま医療センター
******** Click here for English version ********
2016年2月3日 EZR version 1.32公開
2014年11月初心者向けマニュアル刊行
2015年4月EZRマニュアル第2版刊行
フリー統計ソフトEZR(Easy R)とは？

1. 統計解析環境Rとは

それでもやっぱり，プログラミングを頑張ろう!!

- プログラムは，解析の記録．
- 試行錯誤をシステム化．実際の解析では，同じことを何度も繰り返すことになる．
- Rでは，計算結果の出力を柔軟にデザインできる．
- 解析の，実行，記録，保存を自動化できる．

2. Rを使うその前に　データ解析はじめの一歩

データの準備：元データの取り扱い

i. データの形は長方形

- 第一行目に変数名．全角文字は避ける方が無難．
- グラフ、解析結果などを張り付けない．別ファイルで保存．
- データの形は、長方形になるはず．

systemID	hospitalID	sex	age	height	bodyweight
4	1185645	1	64	173	75.4
11	3329388	1	69	164	72
12	4022624	1	78	155.2	47.2
14	4402536	1	83	159.1	60
22	4862866	2	73	147.6	40.5

データの準備：元データの取り扱い（続き）

ii.　元データは絶対に改変しない.

- 解析の過程で、変数を変換したり、新しい変数を定義することがある.
- 新しく作ったデータを、元データに上書きしない.
- データを改変したら、新しいファイル名で保存.
- 元データを改変すると、元データが何であるか分からなくなる．元データが分からなくなれば、意図せざるデータのねつ造まであと一歩.

データの準備：元データの取り扱い（続き）

iii. 患者さんの個人情報は記載しない.

- 残念ながら、いまだに氏名、カルテ番号など、患者さん個人を特定できる情報が付いたままのデータを見かける.
- 個人情報は、データ解析の立場からは無意味.
- 個人情報が漏えいすれば、研究は中止、研究者の辞表が何枚か必要. 被害者には、お詫びの仕様がない.
- データを受け取ったら、個人情報はすぐに匿名化もしくは削除.

データの準備：元データの取り扱い（続き）

iv. 解析記録の保存.

- 患者さんを診察すれば、医師がカルテに記録するのは当然．実験をすれば、実験ノートに記録するのは常識．解析の記録を残すのも、それと同じ.
- 元データと解析の記録を見れば、第三者が解析を再現できる程度の記録が必要.
  - 解析の再現性
  - 備忘録　　「三日後の自分は遠い親戚．一週間後の自分は赤の他人」
  - 出来れば、プログラムを書いて解析する.

データの準備：データ入手時にすべきこと：入力ミス、異常値の発見

Excelのフィルター機能が便利

- データの範囲：本来正の値をとるはずが、負の値をとる．小数点の間違いで、体重35kgが3.5kgになる、等．
- 全角文字と半角文字の混在：“w”と“ｗ” など．
- 質的変数の数字表記：　男性: 1, 女性: 2などを、男性→M, 女性→F のように書き直す．
- 異常な値の検出：“3.14”と“3,14” など．
- 欠測値の数：　欠測値の数が想定より多い場合、データが正常に認識されていないことがある．

記述統計の重要性

- 記述統計はデータを要約し，データの持つ全体的な特徴，傾向を把握する.
- 同じ目的（例：平均の推定）でも，データの持つ性質により複数の解析方法が存在する場合がある. 適切な解析方法を選択するために，データの特徴を把握することが重要.
- データの収集が，公正に行われていることを示す.
  - 比較対照の際，対照のための条件以外の背景因子に，極端な差がないことを示す.
  - データに異常な値がないことを確認.

3. Rを使ったデータ解析の実際

実際のデータ解析は，多くの変数の海の中で意味のある相関を見つけるための試行錯誤.

多くのサブグループ解析と，データの絞り込み.

- データのサブグループ（性別，疾患の有無, etc.）
：データ配列の行方向の切り分け
- 共変量の場合分け：列方向の切り分け
  - 連続変数、離散変数　検定の種類が違う
  - 共変量（データ全体 vs. 術前情報のみ, etc.）
  - 対比の入れ替え（性別，年齢層, etc.）
- アウトカムの場合分け
  - モデルの被説明変数を入れ替える。

3. Rを使ったデータ解析の実際

生存時間解析の例

1. 単変量解析
   - Log-rank test, Cox比例ハザードモデル
   - KM曲線の描画（pngファイルなどで保存）
2. 多変量解析
   - 単変量解析で $p < 0.2$ などの共変量を抽出
   - Cox比例ハザードモデルによる多変量解析
3. 変数選択
   - Backward-eliminationをよく使う。

以上の解析を，アウトカム，サブグループを切り替えて自動的に行う.