

付録 A

R について

A.1 なぜ R を使うべきなのか？

R は R-project (<http://www.r-project.org/>) という国際共同プロジェクトで開発されている，ソースが公開されていて，誰でも自由に利用できる統計ソフトである。ベル研で開発され市販されている S や S-Plus と 8 割くらいは共通の言語仕様をもつとされ，Poorman's S と呼ばれることもある。

R の操作性を市販統計ソフトと比べると，STATISTICA とか JMP とか SPSS のようなメニューから統計手法を選ぶものとはまったく異なり，関数を打たねばならないので，初めて使うときはとっつきにくいかもしれない。が，市販ソフトの中でも，プログラムをタイプするという意味では，SAS と似ている*¹。記述統計から多変量解析まで，およそ必要な統計解析はすべてできるといいし，計算に使われている手法も新しい。大手市販ソフトでは新しい方法のフォローが遅いので，時としてベストではない統計手法が使われたりするのだが，R は現時点でベストとされるかあるいは標準的な手法がデフォルトになっている。たとえば擬似乱数生成のアルゴリズムはバージョン 1.7.0 からメルセンヌツイスター（第 1 章参照）がデフォルトになったし，多重比較のデフォルトは Holm の方法になっている（第 10 章参照）。

プログラムをタイプするのが面倒だと思う人もいるかもしれないが，よく考えてみれば，それが誤解に過ぎないことがわかると思う。少なくとも Microsoft Excel でマクロを使うよりもずっと簡単な場合が多い。

たとえば，Microsoft Excel で独立 2 標本の平均の差の検定をするには，ツールの分析ツール（アドインなので，フルインストールするか，インストール時に指定しな

*¹ もっとも，SAS のプログラムを FORTRAN とすれば，C++ や APL くらいに洗練されている言語体系だと思う。

いと入らない)を使うわけだが、まず等分散性の検定を選び、標本の範囲をそれぞれ指定して実行し、その結果等分散という帰無仮説が棄却されなければ、等分散を仮定した2群の平均の差の検定を選んで再び2つの標本の範囲を選んで実行するし、棄却されたときは等分散でないときの2群の平均の差の検定を選んで2つの標本の範囲を選んで実行する、という手順を踏む必要がある。結果は別々のシートに出力され、それは表として提示できるような形にはなっていない。少なくとも10ステップくらいのマウスの操作が必要であり、わずらわしい。

R ならば、標本サイズが小さければ、変数 x と y (変数名は何でもよい) に直接2つの標本データを付値 (代入) してから、`var.test(x,y)` をして、 p が有意水準未満ならば `t.test(x,y)` でいいし、そうでなければ `t.test(x,y,var.equal=T)` とすればよい。データ範囲を何度も選ぶよりも、 x とか y とかタイプする方が一般的に言えばずっと楽だと思う。もちろん、表形式のデータを読み込んで分析する関数だけ指定することもできる。

しかも、R では、結果を変数に代入して保存したり加工したりできる。`xtable` というライブラリ^{*2}をインストールして読み込めば、結果を `xtable()` の括弧内に入れるだけで、HTML 形式や \LaTeX 形式に変換できたりする。

美しい図を作るのも実に簡単で、しかもその図を PDF とか `postscript` とか `png` とか `jpeg` とか Windows 拡張メタファイル (`emf`) の形式で保存でき、他のソフトに容易に取り込める。たとえば `emf` 形式で保存すれば、Microsoft PowerPoint や OpenOffice.org の Draw などの中で、ベクトルグラフィックスとして再編集できる。PDF 形式で出力してから、 \LaTeX に入っている `pdftops` プログラムで `-eps` オプションをつけて変換すれば、Encapsulated Postscript 形式 (EPS 形式) のファイルを作るのも容易である。

おそらく多くの日本人にとって最大の難点は、日本語が使えない (グラフィック表示は面倒な指定をすればできないこともないし、フォントを変更すればコンソール表示もできるし、データに日本語が入っていてもだいたい扱えるが、変数名としては使えないので、たとえば1行目に日本語を使って変数名を打ってある表データは読み込めない) ことだろう。日本語変数名がなければ、Excel のデータならタブ区切りテキスト形式で保存すれば読みこめるし、`foreign` という標準ライブラリを使って SAS (Transport 形式) や SPSS (.sav 形式) や S-PLUS や Stata や EPIINFO や Octave のデータを読み込むことができる。

日本語による解説があまり出回っていない (英語が読めれば無料でもたくさん出

^{*2} R は拡張が楽なので世界中の研究者が追加のライブラリを作って、R 本体と同じような配布条件で公開しているものが山ほどある。

回っている)のも、多くの日本人にとっては難点かもしれない。統計手法がわかっている、関数名がわからないと実行できないから、統計手法から関数名を探せるようなサービスが欲しいところだろう。本書がその一助になれば幸いである。

A.2 Rを使うための最初の1歩

A.2.1 インストール

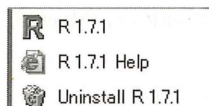
- Rは、現在のところ、Windows, MacOS, Linux, FreeBSDなどのOSの上で利用可能である。
- Windows 版バージョン 1.7.1 のインストールは、CRAN (または会津大学にあるミラー) から `rw1071.exe` をダウンロードして実行するだけなので簡単である。`rw1071.exe` に該当するものは、R-1.6.0 では `rw1060.exe` だったが、それ以前は `SetupR.EXE` というファイル名だったが、いずれにせよインストールの仕方は同じで、ただ実行して、ダイアログに答えていくだけでいい。Windows 版では、デフォルトのインストール先ディレクトリは、`C:\Program Files\R\rw1071` などとバージョン番号が付き、アップグレードしても旧バージョンは自動的に消去されない。なお、手動で追加したライブラリはバージョンアップの際には継承されない (バージョン依存性があるかもしれないから当然だが)、それらのライブラリが zip 形式で公開されているならばそれどこかに保存しておき、R 本体をバージョンアップした後で、`Packages` メニューの `Install package from local zip file` を選んで、保存しておいた追加ライブラリを1つずつ選択するという手順を踏む必要がある^{*3}。
- FreeBSD 4.5R でも、ソースの tar ボールをダウンロードして展開し、そのディレクトリで `./configure` をやってから `make` するだけでコンパイルできるので、`su` して `make install` すればインストールが完了した。Vine Linux 2.6rc1 では `g77` などのフォートランコンパイラを追加インストールする必要があったが、それさえしておけば FreeBSD 4.5R と同様にコンパイルやインストールができた。起動コマンド `R` で起動する `Rconsole` はテキストベースのシェルでも利用できるので、最尤推定などの時間がかかる計算はサーバでやらせると良い。おそらく他の UNIX 系 OS でも似たようなものであろう。

^{*3} ネットワークにつながれた環境であれば、ライブラリは CRAN からダウンロードする方が安全かもしれない。Windows2000/XP, FreeBSD 4.5R, Vine Linux 2.6rc1 の環境では、`install.packages("xtable")` などとするだけで済んで簡単であった。MacOS 版の添付のドキュメントによれば、MacOS 版ではインターネット関連の機能は未実装である。

Debian など、いくつかの Linux ディストリビューションではバイナリパッケージが公開されているので、それをインストールする方が簡単かもしれない。

A.2.2 もっとも基本的な操作

- 起動は、Windows ではデスクトップにできるアイコン（またはスタートメニューのプログラムの R にできるアイコン）をクリックするだけでいい。



Windows2000 なら、コマンドラインでも `Rterm --no-save` として起動できる。Linux や FreeBSD のシェルを telnet や ssh で使う場合は、`R` と打てば良い。いずれの場合でも `>` というプロンプトが表示されて入力待ちになる。

- 終了は、プロンプトに対して `q()` と打てば良い。コマンドラインパラメータとして `--no-save` などとつけて起動した場合以外はワークスペースを保存するかどうかの問い合わせがあるので、その回のセッションを記録しておきたいならば `y` を、そうでなければ `n` と打つ。ワークスペースを保存しすぎると `.RData` というファイルが大きくなって起動が遅くなるが、作業中は便利な機能である。
- 基本的に、関数にデータを与えて得られる結果を表示したり、変数に付値したりして使う。アルファベットとドットからなる文字列は変数になりうる。付値とは、ほぼ代入を意味する。たとえば、`x` という変数に 3, 5, 7 という 3 つの値からなるベクトルを付値するには、`x <- c(3,5,7)` とする。これら 3 つの値の平均を得るには、`mean()` という関数を使って、`mean(x)` とすれば良い。付値せずに関数だけを打てば結果を表示するが、もちろん関数の値を別の変数に付値することもできる。たとえば `y <- mean(x)` として、`y` をまた別の計算に使うこともできる。変数の情報を見るには `str()` という関数が便利である。
- GUI 環境では、Help メニューから R Manual(html) を選べば、階層構造で説明を参照できる。関数へのインデックスもある。
- 関数の使い方を忘れたときは、`help(関数)` とか `help.search("キーワード")` で説明が得られる。
- `example(関数)` で関数の利用例が得られる。
- プログラムは `>` というプロンプトに対して 1 行ずつ入力してもいいが、予

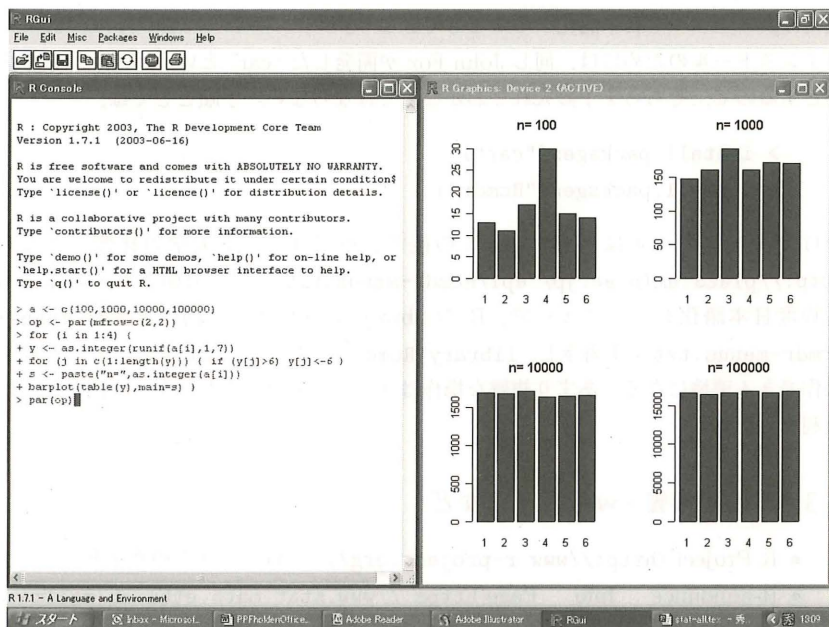


図 A.1 R を Windows で実行した様子。プロンプトは>である。2 行にわたるプログラムを入力中に改行すると、次の行の冒頭に + 記号が自動的に付く。

めテキストファイル（たとえば、d:/test.R）として用意しておけば、`source("d:/test.R")` として一度に読み込んで実行させることもできる。

A.2.3 R Commander を試してみる

sem を初めとして様々なパッケージを開発し、“An R and S-Plus companion to applied regression” という優れた教科書も書いている McMaster 大学の John Fox 教授が、最近になって Rcmdr という、R をメニュー形式で操作するためのパッケージを発表した。

R-1.7.0 以上でないと使えないが、メニュー項目はテキストファイルで定義されているので書き換え可能である。Rcmdr のテキスト表示は tcl/tk で行われており、tcl/tk はバージョン 8.1 から国際化対応しているので、文字コードとして UTF-8 を

使えば、メニューを日本語化することもできる*4。

インストールのためには、同じ John Fox が開発した “car” というパッケージを必要とするので、そちらを予め入れておかななくてはならない。手順としては、

```
> install.packages("car")
> install.packages("Rcmdr")
```

だけでインストールは完了する。この後で、ハーヴァード大学の林啓一さんが <http://plaza.umin.ac.jp/~epi/Rcmdr-menus.txt> として公開されている、ある程度日本語化したメニューで、R の library ディレクトリ内の Rcmdr/menus/Rcmdr-menus.txt を上書きし、library(Rcmdr) とすれば、日本語メニューで R を操作できる環境になる。あまり複雑な操作はメニューではできないが、入門としては便利だと思う。

A.3 R の参考書・web サイトなど

- R Project (<http://www.r-project.org/>): プロジェクトのサイト
- R-announce Info Page(<https://www.stat.math.ethz.ch/mailman/listinfo/r-announce/>): 重要なお知らせが英語で流れる ML の情報ページ。
- CRAN (<http://cran.r-project.org/>): プログラムやライブラリのダウンロード用サイト。
- 会津大学のミラー (<ftp://ftp.u-aizu.ac.jp/pub/lang/R/CRAN/index.html>): CRAN のミラーサイト。日本国内のミラーサイトは、公式にはここだけである。
- 公式入門書やマニュアルの日本語訳は、東京工業大学・間瀬教授のサイト (<http://www.is.titech.ac.jp/~mase/R.html>) で公開されている。その pdf 版は学芸大の森厚さんのサイト (<http://buran.u-gakugei.ac.jp/~mori/LEARN/R/>) からダウンロードできる。なお、2002 年末での最新版及び 2003 年 5 月に公開されたバージョン 1.7.0 の R-intro の暫定和訳については、間瀬教授のサイトからソースファイルをダウンロードし、日本語コードを SJIS に変えて Windows 版 pTeX と dvipdfm を使って個人的に pdf 化したものを、<http://phi.ypu.jp/swtips/R-jp-docs/> にも置いてある（フォントを内蔵していないのでファイルサイズは小さい）。

*4 R-1.7.0 に含まれている tcl/tk はバージョン 8.4 である。

- 群馬大学社会情報学部・青木繁伸教授がRによる統計処理 (<http://aoki2.si.gunma-u.ac.jp/R/>) という凄いいページを作られている。Rを使おうと思う方は必見である。
- 多摩大学・山本義郎助教授によるR-統計解析とグラフィックスの環境 (<http://datamining.tama.ac.jp/~yama/R/>) と R 入門 (<http://datamining.tama.ac.jp/~yama/R/Rintro.html>) は、ちょっと古いバージョンでの解説だが丁寧に書かれていて役に立つと思う。
- Rについての日本語メーリングリスト (R-jp) が、筑波大学のサーバで、岡田昌史さんによって運営されている。元々はドキュメント翻訳用のメーリングリストとしてスタートしたが、Rについて日本語で議論されている ML としては唯一のものである。登録の仕方などの説明は <http://epidemiology.md.tsukuba.ac.jp/~mokada/ml/R-jp.html> にあり、過去に投稿されたメールのアーカイブも公開されている。
- その岡田さんが2003年6月に立ち上げた RjpWiki (<http://www.okada.jp.org/RWiki/>) には、役に立つ情報がものすごい速さで集積しつつある。とくに日本語対応についての進展が著しく、日本語変数名を使えるようにするパッチなども発表されている。
- 山口県立大学の中澤のサイト内でも <http://phi.ypu.jp/swtips/R.html> として R についての使い方の tips や最新情報を公開している。本書のサポートもここで行う予定である。応用例として、「Rでシミュレーション」(<http://phi.ypu.jp/swtips/Rsim/>)や、「人口ピラミッドの作り方」(<http://phi.ypu.jp/demography/makepyramid.html>) も参考にされたい。