パッケージの開発・登録とメンテナンスについて

長島 健悟

城西大学 薬学部 助手

Kengo NAGASHIMA

Faculty of Pharmaceutical Sciences, Josai University

2010 年度 統計数理研究所共同研究集会
「データ解析環境 R の整備と利用」
2010 年 11 月 27 日

自己紹介

- 職業は城西大学薬学部の助手
  仕事は教育及び研究
- 専門領域
  - 医薬統計学 (Biostatistics)
  - ゲノム疫学

- R の利用場面
  - 試験計画のための事前準備 (サンプルサイズ設計, シミュレーション)
    試験結果の解析／グラフの作成
  - 新手法の実装

R パッケージ開発の話をする理由

- 新しい統計手法の再現性を求める国際的な流れ
- その結果の利用が適切に行われているか

生物統計学研究における動き

- 一部の国際学会誌で再現性 (Reproducibility) のある研究が求められてきている
  - 計算プログラム, データの公開を推奨

- Biostatistics (Peng 2009)[1]
  Data: ...
  Code: Any computer code, software, or other computer instructions that were used to compute published results are provided. For software that is widely available from central repositories (e.g. CRAN, Statlib), a reference to where they can be obtained will suffice.
  Reproducible: ...
- Biometrical Journal (Aims and Scope)[2]
  The Editors are supporting reproducible research. Authors are strongly encouraged to submit computer code and data sets used to illustrate new methods.

ゲノムデータ解析研究における動き

- International Society for Computational Biology のソフトウェア共有の声明 [4]

  This software sharing statement is intended to address the ability of the scientific community to reproduce and build on research findings reported in scientific publications or generated with public funds.

  ...
- Bioconductor[3]

  ハイスループットなゲノムデータを解析し, 理解するためのオープンソースソフトウェア. たくさんの R パッケージから構成されている.

  The broad goals of the Bioconductor project are:

  ...

  To further scientific understanding by producing high-quality documentation and reproducible research.

  ...

R は使いやすいか？

- 研究環境としては使いやすい
  世界的な動きもあり, 継続的に発展すると思われる
- データ解析には使いにくい
  日本語対応が不十分
  ある程度スクリプトが書けなくてはいけない
- Brian Ripley 教授の言葉 [5]
  “Let’s not kid ourselves: the most widely used piece of software for statistics is Excel.”
- 研究の統計解析について相談を受けると, ほとんどが Excel のアドオン利用者 (薬学部)
  だが, Excel は欠陥がある上になかなか修正されない [6, 8, 7]

新しい方法があまり利用されず
もったいない

本発表の目的

- 研究結果の利用可能性を高めるために

Rパッケージ開発の流れと
多言語化(日本語対応)の方法を概略する

パッケージの作成

本日の発表について

- R のバージョンは 2.10.1 (Rtools のバージョンは 2.10) を使用して資料を作成しています
- Windows 環境での場合を扱います
  - Linux または MacOS X 環境については, 適宜読み替え／互換品をご利用ください

パッケージ作成に関する資料

これを読むと完璧

- Writing R Extensions
  http://cran.r-project.org/doc/manuals/R-exts.pdf

Web 上に日本語の資料もいくつかある

- Writing R Extensions (日本語訳, 2001 年のもの)
  http://cran.r-project.org/doc/contrib/manuals-jp/R-exts.jp.pdf
- 私的パッケージ作成法 - RjpWiki
  http://www.okada.jp.org/RWiki/?%BB%E4%C5%AA%A5%D1%A5%C3%A5%B1%A1%BC%A5%B8%BA%EE%C0%AE%CB%A1
- 10 分で分かる R パッケージの作り方
  http://www.slideshare.net/yokkuns/10r

R パッケージの一番簡単な作り方

1. 統計処理を行う関数やデータを作成する
2. 関数 package.skeleton() を実行して雛形を作成する
3. 雛形を適切な形に加工する
   DESCRIPTION, NAMESPACE, Rd ファイル
4. 誤りがないかチェックを行う

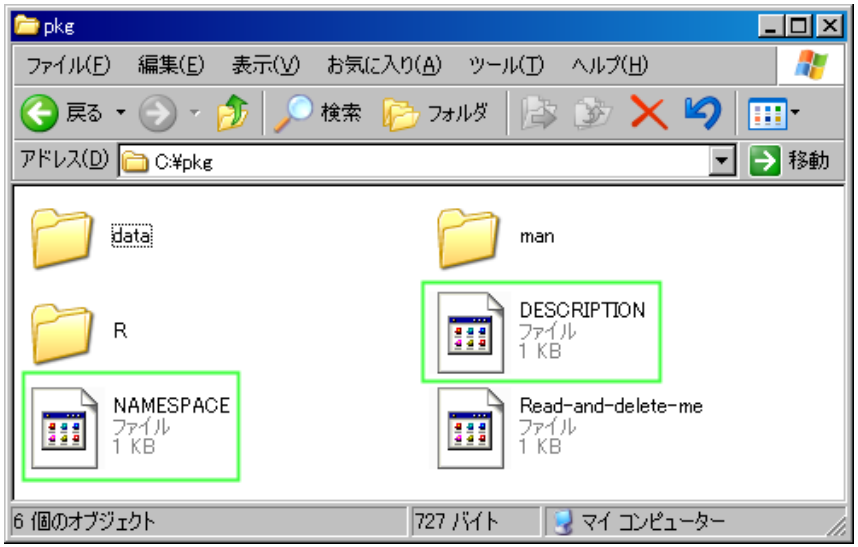
package.skeleton(): 関数からパッケージ雛形を作る (1)

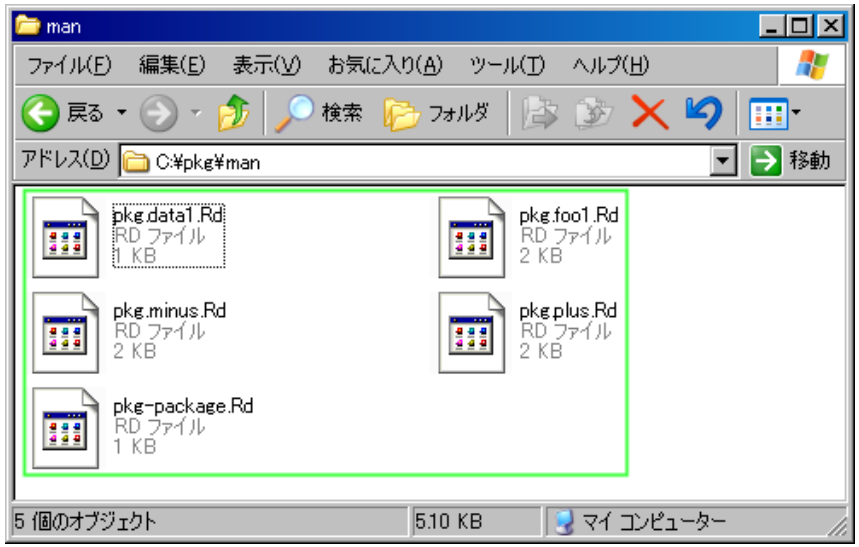
```
library(tools); require(stats)
pkg.data1 <- list(a=rnorm(2), b=rnorm(2))
pkg.plus  <- function(a, b) { return(a + b) }
pkg.minus <- function(a, b) { return(a - b) }
pkg.foo1  <- function(x, y) { return(rnorm(1,x)*rnorm(1,y)) }
```

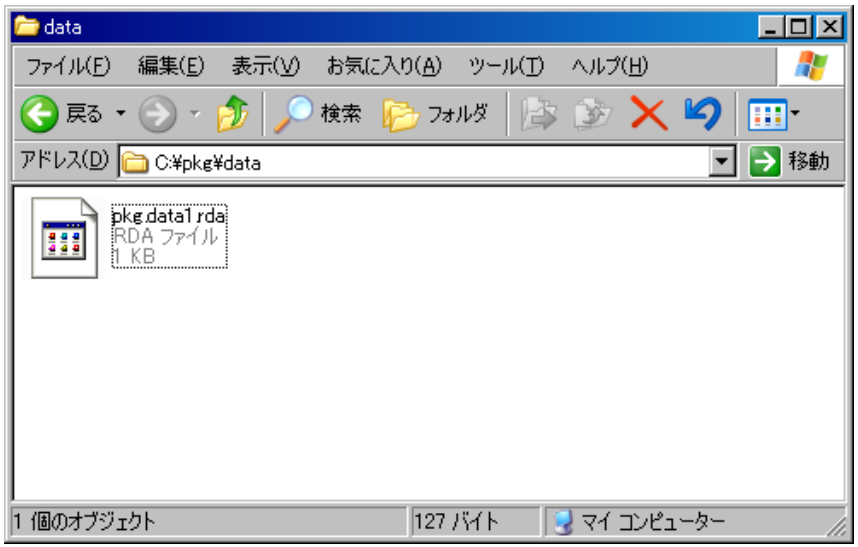
```
package.skeleton(
  name="pkg",
  list=c("pkg.data1", "pkg.plus", "pkg.minus", "pkg.foo1"),
  path=".", force = TRUE, namespace = TRUE)
```

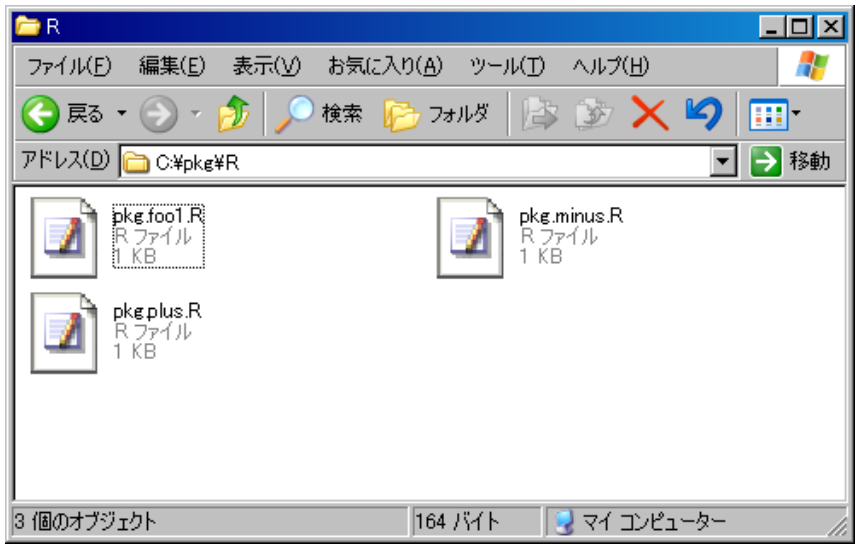
- "path/pkg/" に必要なファイルと, 一部の雛形を生成

package.skeleton(): 関数からパッケージ雛形を作る (2)







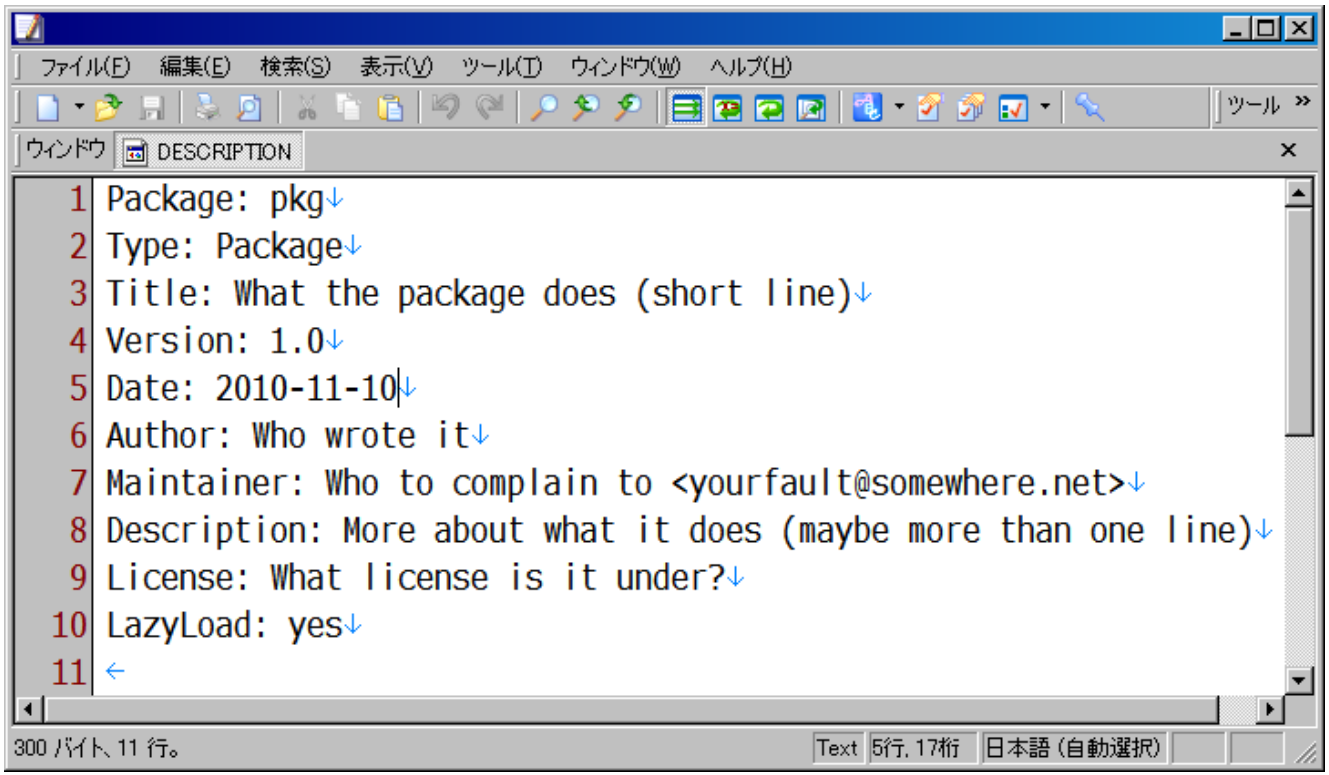


生成されるファイル

- /DESCRIPTION
  パッケージの説明ファイル
- /NAMESPACE
  名前空間の設定ファイル
- /man/(パッケージ名)-package.Rd, /man/(関数名 or データ).Rd
  パッケージ or 関数 or データのドキュメントファイル
- /R/(関数名).R, /data/(データ名).R
  関数ファイルとデータファイル
- 他にも目的に応じて, 様々なファイルが追加できる
  Writing R Extensions - Package structure

/DESCRIPTION: パッケージの説明ファイル

- パッケージの情報
  'Package', 'Version', 'License', 'Description', 'Title', 'Author', 'Maintainer' は必須の項目



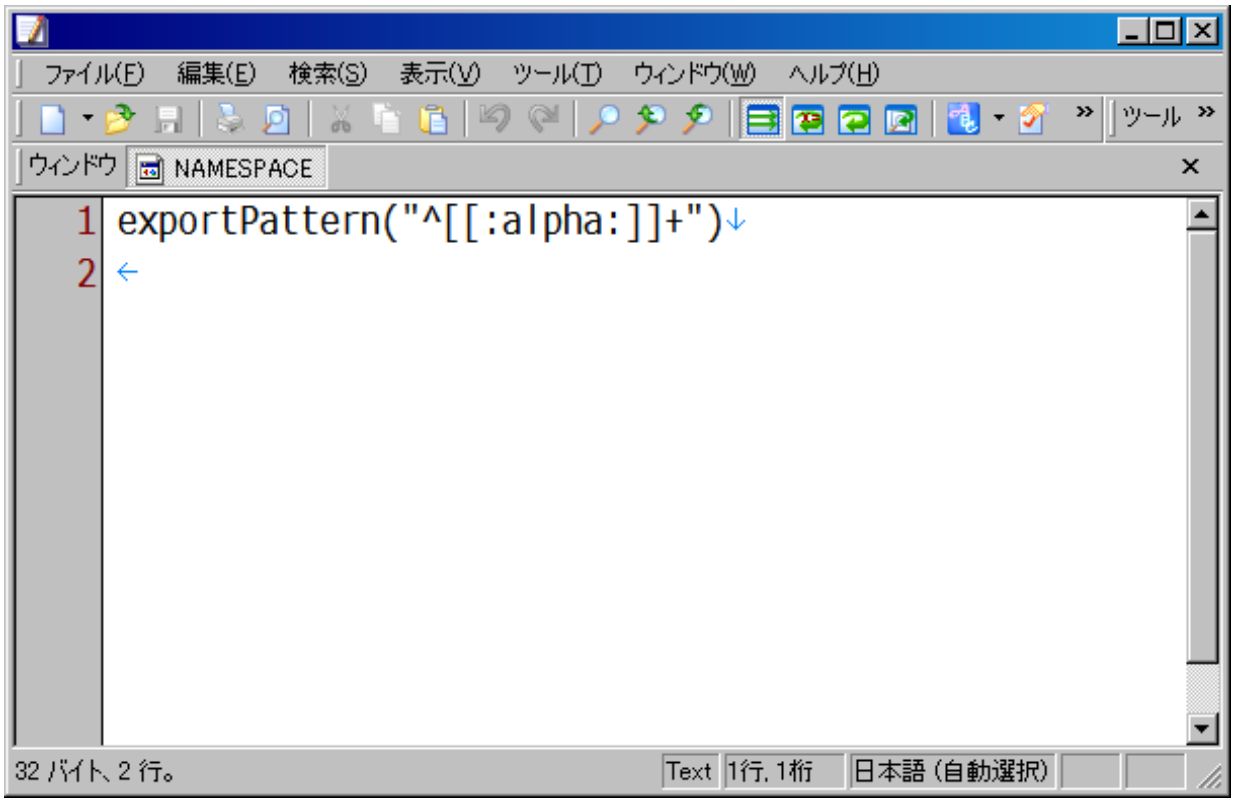
/DESCRIPTION: パッケージの説明ファイル

```
Package: pkg
Title: Sample Package (pkg)
Version: 1.0-1
Author: Kengo NAGASHIMA
Maintainer: Kengo NAGASHIMA <nagasima@josai.ac.jp>
Description: This package is sample package, which contain useless
    sample code.
License: GPL (>= 2)
Depends: R (>= 2.10.1), tools, stats
```

- 'Version' の記法
  '0.01', '0.01.0', '0.1-0' など
- 'License' の記法
  LGPL (>= 2.0, < 3) | Mozilla Public License, GPL-2 | file LICENCE など
- 'Maintainer'
  一名のみでメールアドレスを併記
- その他の情報 (パッケージの依存関係なども指定する)
  Writing R Extensions - The DESCRIPTION file

/NAMESPACE: 名前空間の設定ファイル

- package.skeleton(..., namespace=TRUE) を指定していれば, 全てのオブジェクトが指定される
  データ pkg.data1, 関数 pkg.plus, pkg.minus, pkg.foo1 が外部から読み込みできる



/NAMESPACE: 名前空間の設定ファイル

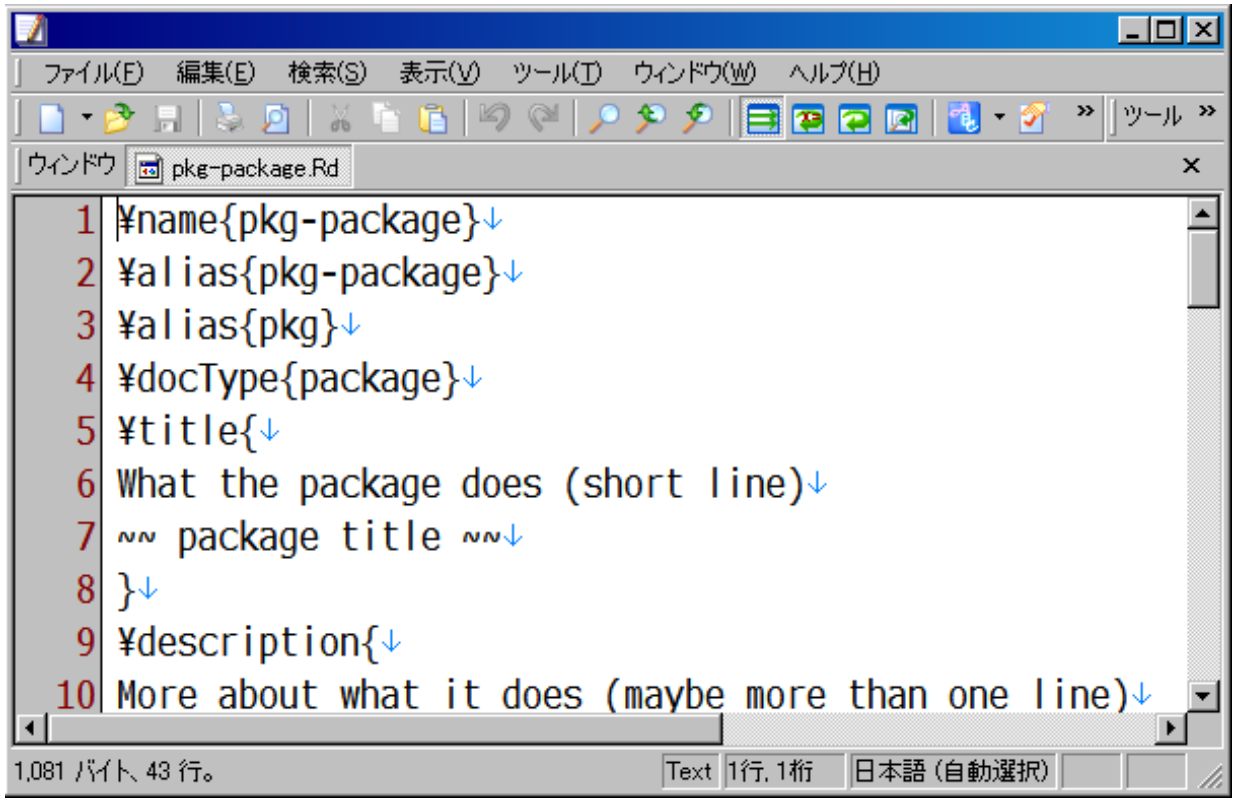
- 外部から呼び出しできるオブジェクトの指定
  export(function1, data1)
- exportPattern() は指定した正規表現にマッチするオブジェクトをエクスポートする

```
exportPattern("^[[:alpha:]]+")
```

- 外部パッケージから読み込むオブジェクトの指定
  import(pkg, function2, data2)
- S3 メソッドの指定
  S3method(print, function3)
- C や Fortran で書かれたライブラリの指定
  useDynLib(library)
- その他の詳しい記述
  Writing R Extensions - Package name spaces

/man/*.Rd: ドキュメントファイル

- pkg-package.Rd
  pkg.data1.Rd
  pkg.plus.Rd, pkg.minus.Rd, pkg.foo1.Rd
- “Rd format” と呼ばれる形式で記述する



/man/*.Rd: ドキュメントファイル

- `\name{name}`
  ドキュメント名
- `\alias{topic}`
  関連トピックを指定する, ヘルプの検索に利用される
- `\title{Title}`
  ドキュメントのタイトル
- `\description{...}`
  内容を数行でまとめて記述する
- `\usage{fun(arg1, arg2, ...)}`
  簡単な使い方を記述する
- `\arguments{...}`
  引数の説明
- `\details{...}`
  description で書けなかった手法の細かい内容などを記述する

/man/*.Rd: ドキュメントファイル

- `\value{...}`
  返り値の説明
- `\references{...}`
  引用・参照する文献
- `\note{...}`
  他に特別に言及しておきたいことがあれば記述する
- `\author{...}`
  Author と Maintainer を記述する, メールアドレスは`\email{}`, URL は`\url{}`で記述できる
- `\seealso{...}`
  他に参照する必要がある関数やパッケージなどを記述する, パッケージ内リンクは Writing R Extensions - Marking text, 外部パッケージへのリンクは Writing R Extensions - Cross-references

/man/*.Rd: ドキュメントファイル

- `\examples{...}`
  実行可能なサンプル (R のスクリプト) を記述する
- `\keyword{key}`
  関連キーワードを指定する, 'R_HOME/doc/KEYWORDS' というファイル内のリストからキーワードを選ぶ
- その他詳細
  Writing R Extensions - Writing R documentation files

/man/*.Rd: ドキュメントファイル

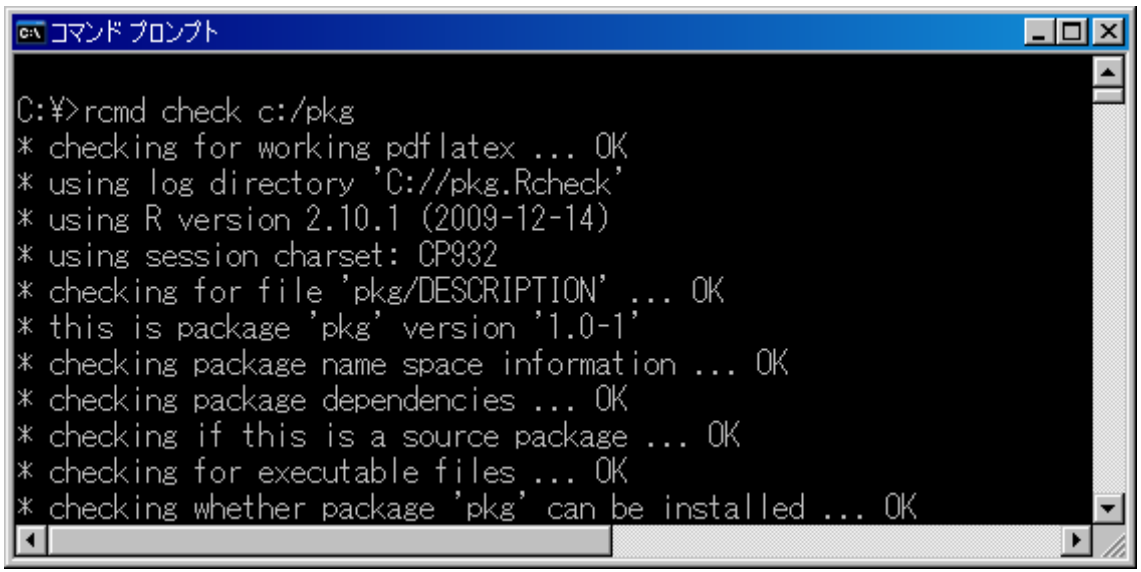
```
\name{pkg-package}
\alias{pkg-package}
\alias{pkg}
\docType{package}
\title{
Sample Package (pkg)
}
\description{
This package is sample package, which contain useless sample code.
}
\details{
\tabular{ll}{
Package: \tab pkg\cr
Version: \tab 1.0-1\cr
License: \tab GPL (>= 2)\cr
Depends: \tab R (>= 2.10.1), tools, stats\cr
}
}
\author{
Kengo NAGASHIMA

Maintainer: Kengo NAGASHIMA <nagasima@josai.ac.jp>
}
\keyword{ package }
\seealso{
\code{\link{pkg.data1}}, \code{\link{pkg.plus}}, \code{\link{pkg.minus}}, \code{\link{pkg.
      foo1}}
}
\examples{
pkg.plus(1, 2)
}
```

rcmd check: パッケージのチェック

- コマンドプロンプトなどで, rcmd check を実行する

```
rcmd check path/calc
```

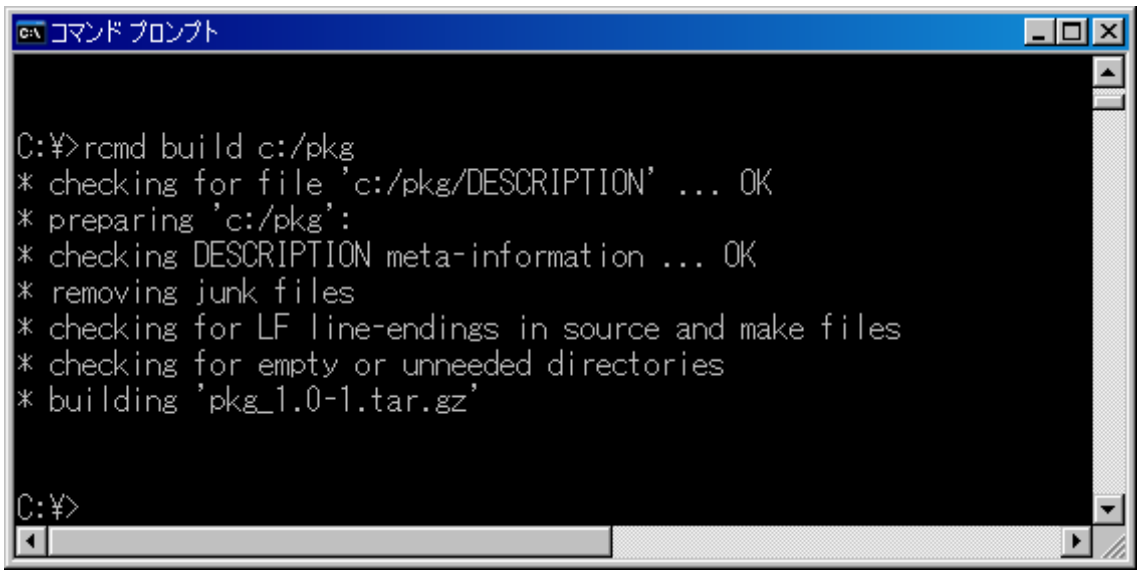


- NG (や WARNING) が出なくなるまで問題点を修正し, 全てのチェックが通ったらアップロード可能な状態
  Writing R Extensions - Checking packages

アップロードの準備 - パッケージのビルド

- コマンドプロンプトなどで, rcmd build を実行する (‘.tar.gz’ ファイルを作成)

```
rcmd build path/calc
```



- 他の注意点など
  Writing R Extensions - Building packages

アップロードと報告 (1)

- 作成したパッケージの整合性をチェックし, エラー等が起きないことを確認できたら, CRAN サーバに登録します
- CRAN の ftp サーバ (ftp://CRAN.R-project.org/incoming/) に, '.tar.gz' ファイルをアップロードする (他の形式は不可)
- アップロードしたらメールで報告して, 返信を待ちます

```
Subject: Report of uploading (mmcm_1.1-0.tar.gz)

Hello CRAN coordinator.

I was uploading the update of "Modified Maximum Contrast Method"
package (mmcm_1.1-0.tar.gz) to R-project ftp server. According
your request, I send you the confirmation e-Mail now.

Please inspect my uploading contents.

Sincerely yours,
--
Kengo NAGASHIMA
```

アップロードと報告 (2)

- アップロードしたパッケージがチェックされ, 問題が無ければ, 登録作業開始の報告が来ます

```
Subject: Package mmcm_1.1-0.tar.gz has been built for Windows

Dear package maintainer,

this notification has been generated automatically.
Your package mmcm_1.1-0.tar.gz has been built for Windows and will
be published within 24 hours in the corresponding CRAN directory
(CRAN/bin/windows/contrib/2.10/).
R version 2.10.1 (2009-12-14)
```

- 上記のメールを受信後, Package source と Windows binary は 1 日程度で登録され, MacOS X binary は数日かかっていた (2009 年頃の情報)

メンテナンス

- パッケージのアップデートや修正の必要が生じた場合, 修正したファイル再度アップロードして報告を行うだけでよい
- 基本的な手順は同じで, パッケージの修正, パッケージのチェック, パッケージのビルド, '.tar.gz' ファイルのアップロード, メールによる報告となる

- 64bit 対応になったときに, 修正依頼のメールを 1 度いただいたが, 他に連絡をもらったことはない

多言語化

多言語化の基本

- R 側の対応
  gettext() 関数
- 翻訳語の記述
  po ファイル

gettext() 関数

- gettext("翻訳対象文字列", domain="R-(パッケージ名)") 関数
  翻訳対象となる文字列をあらかじめ指定しておくための関数
- gettextf("翻訳対象文字列", ..., domain="R-(パッケージ名)") 関数
  C 言語の printf() 関数の様にフォーマット指定できる
- 例えば, pkg.plus 関数を以下の様に書き換えるとします

```
pkg.plus <-
function(a, b) {
 msg <- paste(
  gettext("This is pkg.plus() function.", domain="R-pkg"), "\n",
  gettextf("a = %f, b = %f.", a, b, domain="R-pkg"), "\n",
  sep=""
 )
 cat(msg)
 return(a + b)
}
```

- R 側の準備はこれで終わり

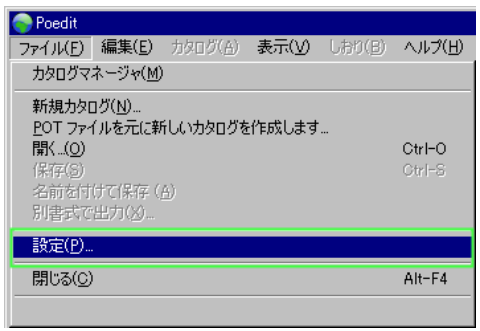
po ファイル

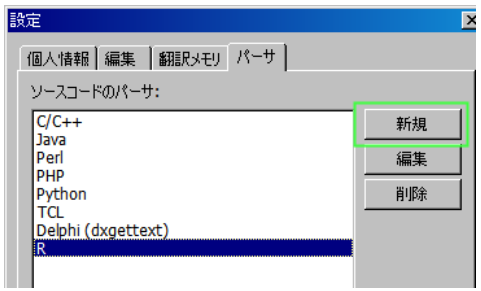
- ディレクトリ "path/inst/po/(言語名)/LC_MESSAGES" を追加
- パッケージ用ファイルのディレクトリ構成が以下の様になっているか確認

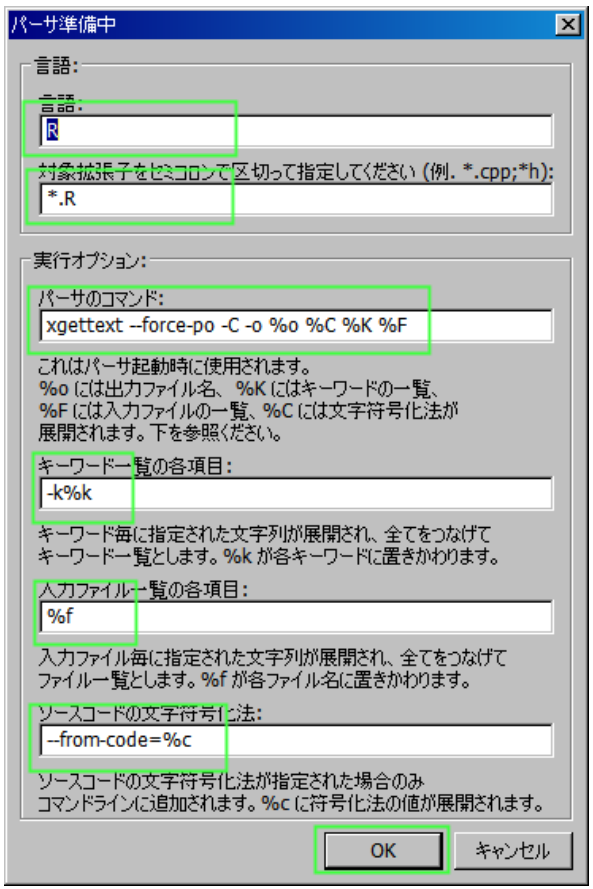
- "path/inst/po/(言語名)/LC_MESSAGES" に mo ファイルを配置することで多言語化される
- po ファイルは mo ファイルを作成するためのファイル
- Poedit というソフトウェアを利用できる

Poedit の設定

- パーサの設定
  コマンドを “xgettext –force-po -C -o %o %C %K %F” に変更し, 後は C 言語等からコピーすればよい

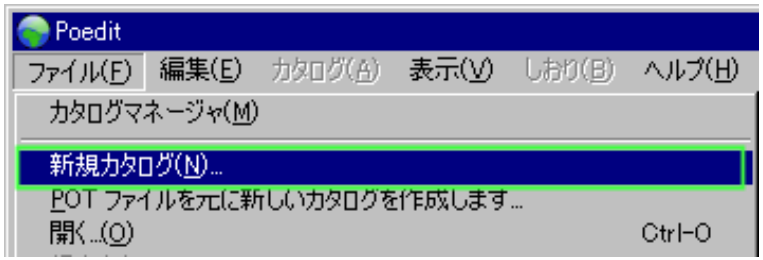


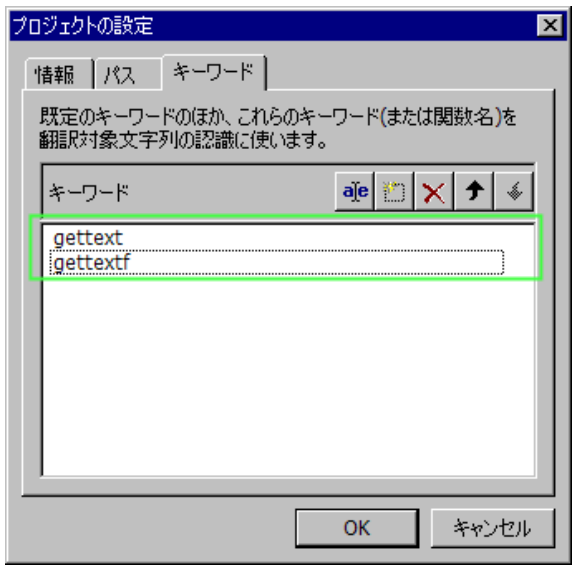




po ファイルの作成 (1)

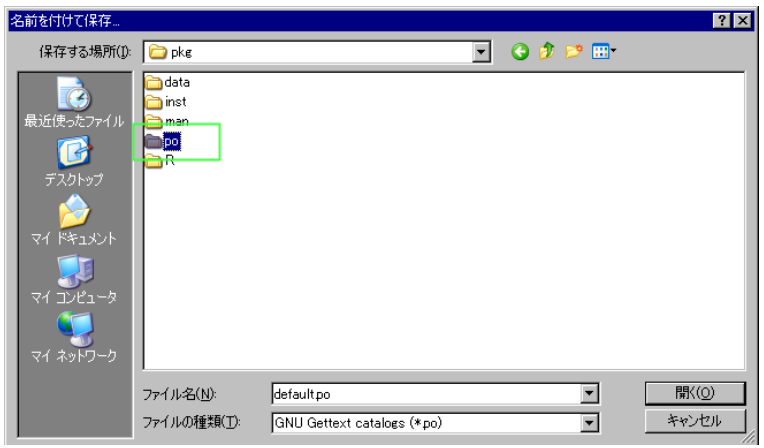
- 新規カタログを作成
- プロジェクト情報を埋め, パスの設定でベースパスを “../”, パスに “./R/” を追加, キーワードを “gettext”, “gettextf” に変更

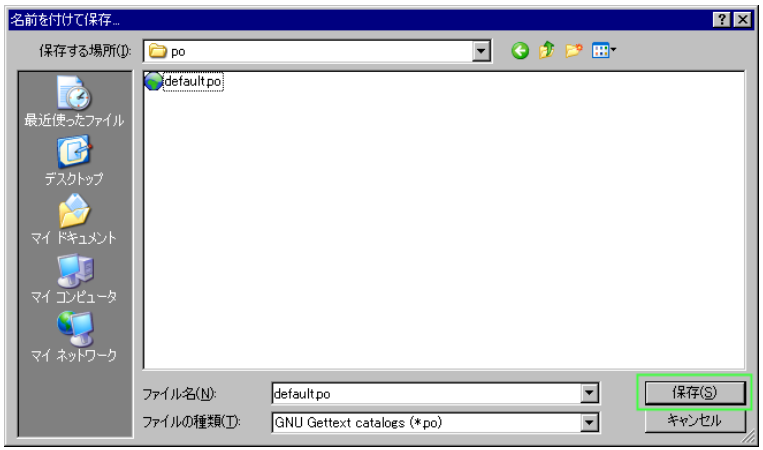




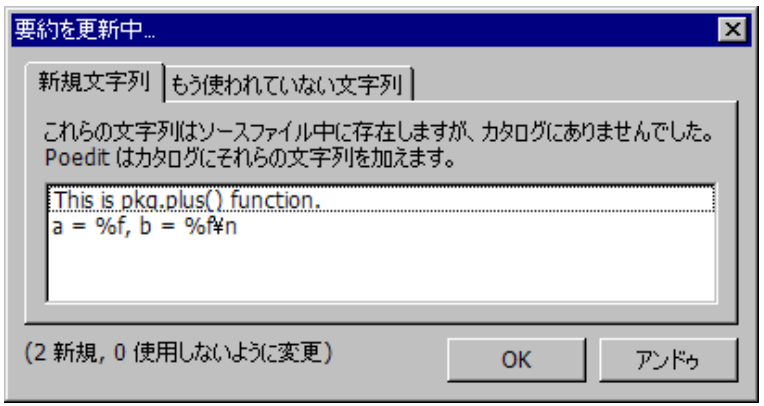
po ファイルの作成 (2)

- 保存先を "path/pkg/po" に指定する



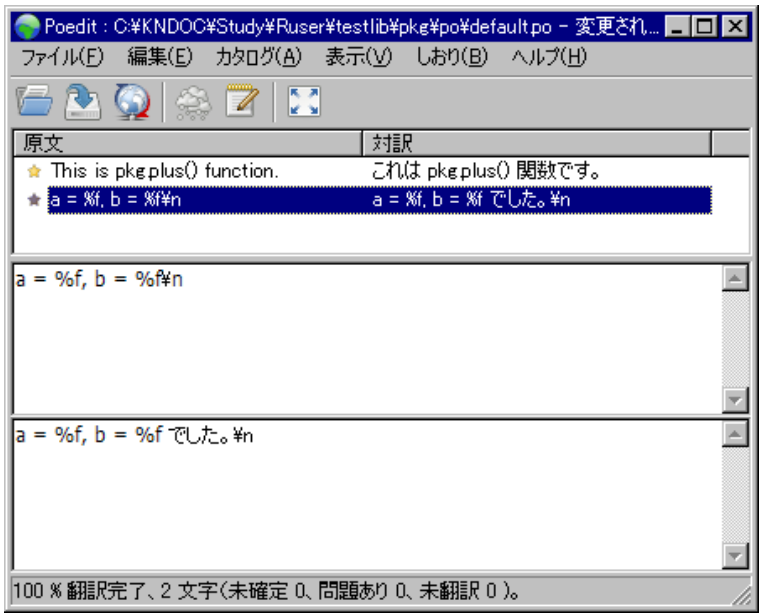


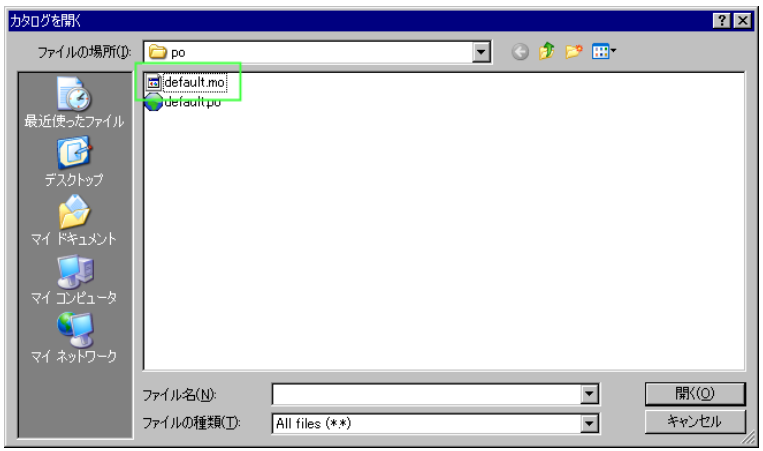
- 翻訳対象文字列の抽出処理が完了する



翻訳作業と mo ファイルの生成

- 下の窓に翻訳したメッセージを入力し, 保存すると自動的に mo ファイルが生成される





- mo ファイルの名前を “R-pkg.mo” にリネーム
- 最後に "path/inst/po/(言語名)/LC_MESSAGES" に作成された mo ファイルをコピーして終了

パッケージファイルの作成と結果の確認

- Windows 用バイナリを作成し, R にインストールして言語を切り替えながら結果を確認してみる

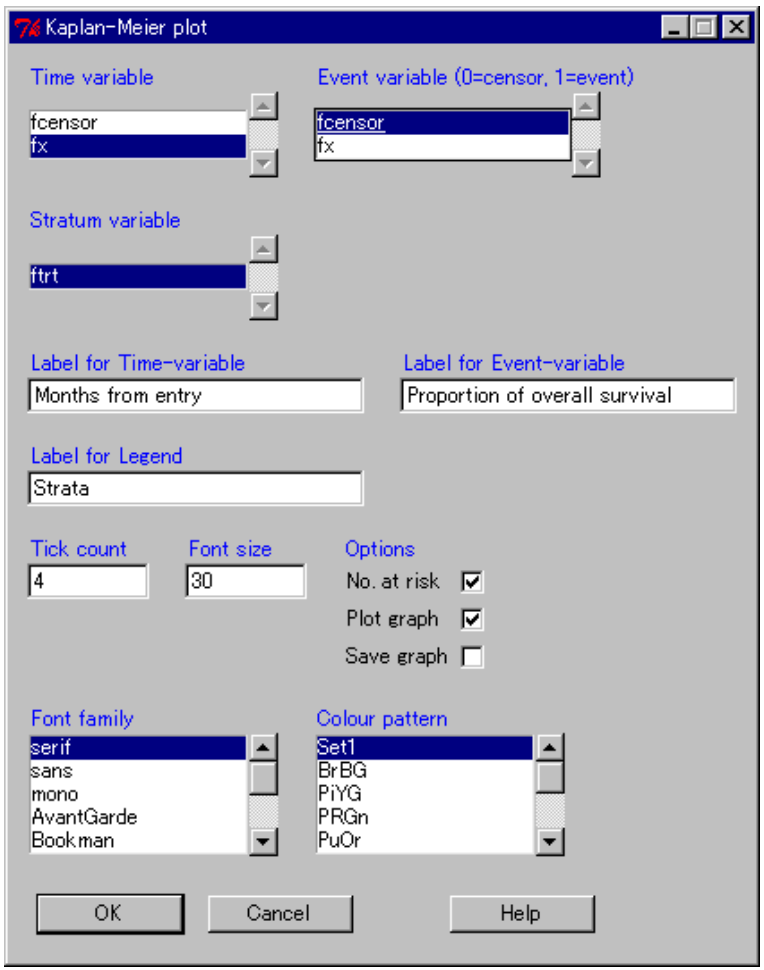
```
rcmd build --binary path/pkg
```

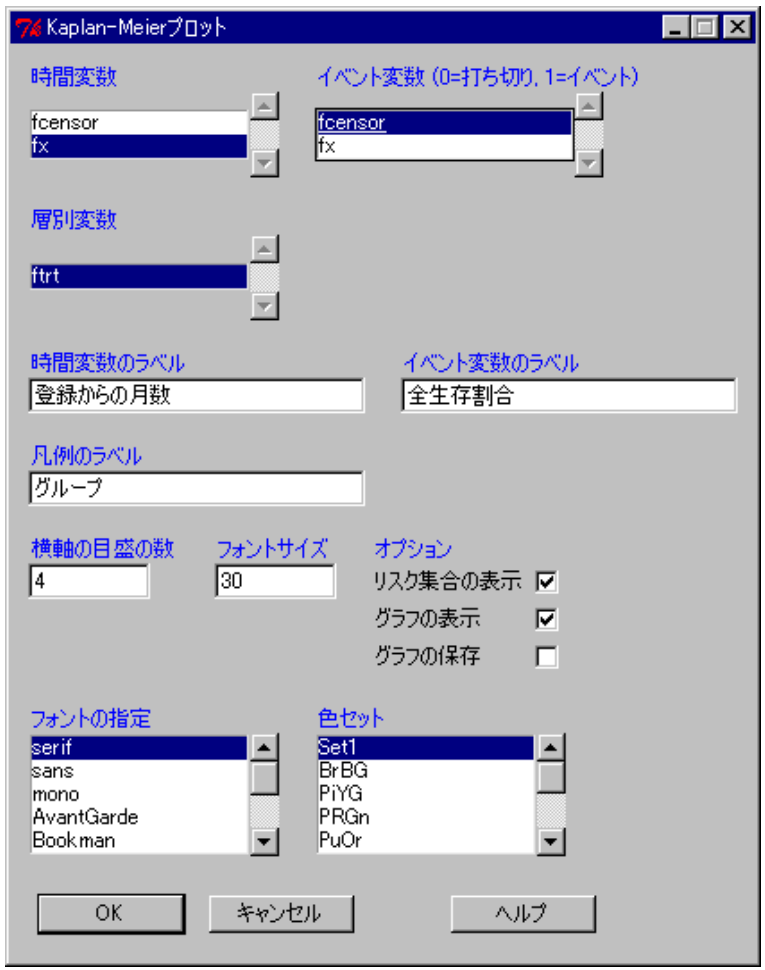
```
> install.packages(paste(path , "pkg_1.0-1.zip", sep=""),
 パッケージ 'pkg' は無事に開封され、MD5 サムもチェックされ
>
> library(pkg)
 要求されたパッケージ tools をロード中です
> pkg.plus(1,2)
これは pkg.plus() 関数です。
a = 1.000000, b = 2.000000 でした。
[1] 3
>
```

```
> install.packages(paste(path , "pkg_1.0-1.zip", sep=""),
package 'pkg' successfully unpacked and MD5 sums checked
>
> library(pkg)
Loading required package: tools
> pkg.plus(1,2)
This is pkg.plus() function.
a = 1.000000, b = 2.000000.
[1] 3
>
```

最近作成中のパッケージ

- RcmdrPlugin.KMggplot2: ggplot2 を Rcommander から操作できるプラグイン

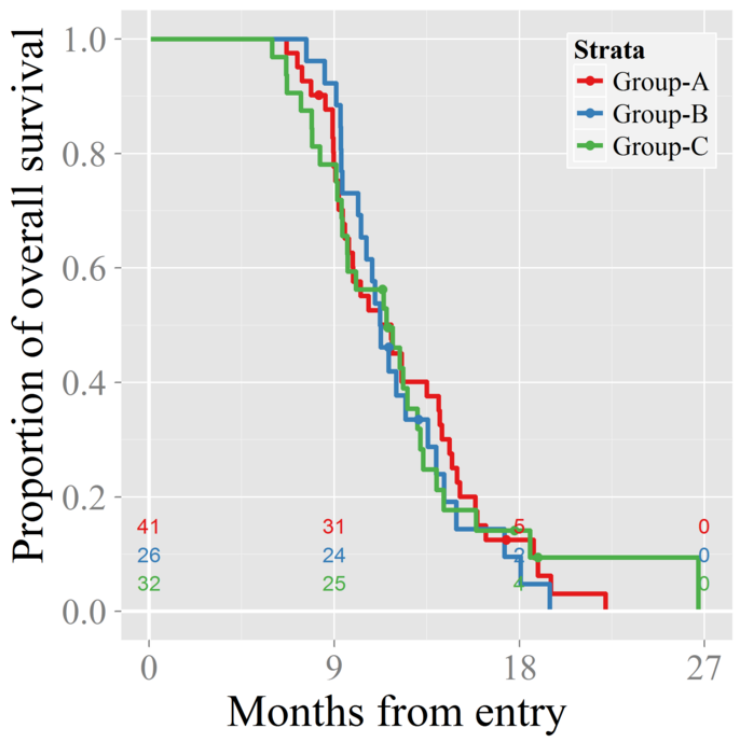




RcmdrPlugin 作成時の注意点

- Rcmdr パッケージには, 多言語に対応するための gettextRcmdr() という関数があるが, これは使えなかった
- gettext() 関数を domain をきちんと指定して使う必要があった

```
gettext(..., domain="R-RcmdrPlugin.KMggplot2")
```



R ファイル側

```
‘kmg2.button‘ <- function() {

  ‘kmg2.gettextRcmdr‘ <- function(...) {
    gettext(..., domain="R-RcmdrPlugin.KMggplot2")
  }

  initializeDialog(title=kmg2.gettextRcmdr("Kaplan-Meier plot"))

  variablesFrame <- tkframe(top)
  xBox <- variableListBox(variablesFrame, Numeric(), title=kmg2.
      gettextRcmdr("Time variable"),
    listHeight=5, selectmode="single")
  yBox <- variableListBox(variablesFrame, Numeric(), title=kmg2.
      gettextRcmdr("Event variable (0=censor, 1=event)"),
    listHeight=5, selectmode="single")
  zBox <- variableListBox(variablesFrame, Factors(), title=kmg2.
      gettextRcmdr("Stratum variable"),
    listHeight=5, selectmode="single", initialSelection=0)
```

po ファイル側

```
msgid "Kaplan-Meier plot..."
msgstr "Kaplan-Meierプロット..."

msgid "Kaplan-Meier plot"
msgstr "Kaplan-Meierプロット"

msgid "Time variable"
msgstr "時間変数"

msgid "Event variable (0=censor, 1=event)"
msgstr "イベント変数 (0=打ち切り, 1=イベント)"

msgid "Stratum variable"
msgstr "層別変数"
```

その他の注意点

- R のみで記述されたパッケージであれば, 今回紹介したパッケージ作成・多言語化の流れはほとんど変わりません
- C/C++や FORTRAN を利用したり, OS に依存する処理を行う場合は, Writing R Extensions などを参考にします

最後に

- よりよいパッケージ, より使いやすいパッケージを作成し, 日本の統計コミュニティが発展していくことを望んで活動しています
- この資料はリンク付きスライドですので, 後日資料ファイルをホームページ上で公開予定です http://www.josai.ac.jp/~nagasima/

参考文献

[1] Peng RD. Reproducible research and Biostatistics. *Biostatistics* 2009; **10**(3): 405–408.

[2] Biometrical Journal, Overview, Aims and Scope. [cited 2010 Sep. 27]; Available from: http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1521-4036/homepage/ProductInformation.html

[3] Bioconductor, About Bioconductor. [cited 2010 Nov. 11]; Available from: http://www.bioconductor.org/about/index.html

[4] International Society for Computational Biology, Software Sharing Policy Statement. [cited 2010 Nov. 10]; Available from: http://www.iscb.org/iscb-policy-statements-/187

[5] Ripley BD. Statistical methods *need* software: A view of statistical computing. Presentation RSS Meeting, September 2002. [cited 2010 Nov. 11]; Available from: http://www.stats.ox.ac.uk/~ripley/RSS2002.pdf

[6] McCullough BD, Heiser DA. On the accuracy of statistical procedures in Microsoft Excel 2007. *Computational Statistics & Data Analysis* 2008; **52**(10): 4570–4578.

[7] Yalta TA. The accuracy of statistical distributions in Microsoft®Excel 2007. *Computational Statistics & Data Analysis* 2008; **52**(10): 4579–4586.

[8] McCullough BD. Microsoft Excel's 'Not The Wichmann–Hill' random number generators. *Computational Statistics & Data Analysis* 2008; **52**(10): 4587–4593.

[9] CRAN. Writing R Extensions. [cited 2010 Nov. 11]; Available from: http://cran.r-project.org/doc/manuals/R-exts.html