R を用いた人口分析の教育事例：長所と短所

中澤　港（群馬大学）

2010 年 6 月 14 日

1　R の基本

R は誰でも自由に使えるソフトである。Windows でも Mac OS でも Linux でも動作する。インストールすべきファイルは，CRAN (The Comprehensive R Archive Network) からダウンロードすることができる。CRAN のミラーサイトが各国に存在し，ダウンロードは国内のミラーサイトからすることが推奨されている。日本では

- 会津大学：`ftp://ftp.u-aizu.ac.jp/pub/lang/R/CRAN/`
- 筑波大学：`http://cran.md.tsukuba.ac.jp/`
- 東京大学：`http://ftp.ecc.u-tokyo.ac.jp/CRAN/`

のどれかを利用すべきだろう。

2010 年 6 月 12 日現在の最新安定版は 2.11.1 なので，Windows ならば `R-2.11.1-win32.exe` をダウンロードし，ダブルクリックして実行するとインストールが始まる。インストーラが UNICODE 版なのに日本語メッセージの一部が UNICODE になっていないためか，Japanese でインストールをするとカスタムインストールの選択肢が文字化けする不具合があり，インストールに使用する言語は English を選ぶべきである。

インストールが完了すれば，R の起動は，デスクトップか Quick Launch かスタートメニューにある起動アイコンを選ぶだけでいい。自動的に赤い > 記号が表示されて入力待ちになる。この記号>をプロンプトと呼ぶ。R への対話的コマンド入力はプロンプトに対して行う。ただし，「ファイル」の「新しいスクリプト」か「スクリプトを開く」でスクリプトエディタを開いてコードを編集し，一括実行させる方が便利である。閉じ括弧を付け忘れたり命令や関数の途中で改行してしまった場合は（関数文字列の途中ではエラーになるが，ちょうど切りが良ければ）プロンプトが継続行を意味する + となるので，本来打つつもりだった文字列を続けて入力していい。なお，Windows では，どうしても継続行状態から抜けられなくなってしまった場合，ESC キーを押すとプロンプトに戻ることができる。

入力した命令や関数は，「ファイル」メニューの「履歴の保存」で保存でき，後で「ファイル」メニューの「履歴の読み込み」または「R コードのソースを読みこみ」で呼び出せば完全に再現できる。プロンプトに対して source("プログラムファイル名") としても同じことになる。なお，Windows ではファイルパス中，ディレクトリ（フォルダ）の区切りは/または￥￥で表す。できるだけ 1 つの作業ディレクトリを決めて作業することにする方が簡単である。作業ディレクトリは起動アイコンのプロパティの「作業フォルダ」で指定できる。また，キーボードの↑を押せば既に入力したコマンドを呼び戻すことができる。

1.1　基本コマンド

終了　`q()`

付値　`<-` 例えば，1，4，6 という 3 つの数値からなるベクトルを `x` という変数に保存するには次のようにする。

```
x <- c(1,4,6)
```

定義　`function()` 例えば，平均と標準偏差を計算する関数 `meansd()` の定義は次の通り。

```
meansd <- function(X) { list(mean(X),sd(X)) }
```

導入　`install.packages()` 例えば，CRAN から `Epi` パッケージ[*1]をダウンロードしてインストールするには[*2]，

```
install.packages("Epi",dep=TRUE)
```

とする。最初のダウンロード利用時には，ライブラリをどのミラーサーバからダウンロードするかを聞いてくるので国内のミラーサーバを指定する。発表者は筑波大学のサーバを利用することが多い。`dep=TRUE` は dependency（依存）が真という意味で，`Epi` パッケージが依存している `Epi` 以外のパッケージも自動的にダウンロードしてインストールしてくれる。なお，`TRUE` は `T` でも有効だが，誤って `T` を変数として別の値を付値してしまっていると，意図しない動作をしてしまい，原因を見つけにくいバグの元になるので，できるだけ `TRUE` とフルスペル書いておく

[*1] `http://staff.pubhealth.ku.dk/~bxc/Epi/`に詳細な説明があるが，デンマーク・コペンハーゲン大学の Bendix Carstensen らが開発して CRAN で公開している，慢性疾患の疫学のためのライブラリである。人口分析関係では age-period-cohort モデルの当てはめや Lexis diagram を描く関数が含まれている。

[*2] Windows Vista または 7 の場合は，管理者権限がないとパッケージのインストールができないし，他にもいろいろな制約がある。

ことが推奨されている。他に依存するパッケージがないパッケージ，例えば発表者が開発して公開してある pyramid パッケージをインストールする場合は，install.packages("pyramid") とするだけでよい。

ヘルプ　？例えば，t 検定の関数 t.test の解説をみるには，?t.test とする。

関数定義は何行にも渡って行うことができ，最終行の値が戻り値となる。関数内の変数は局所化されているので，関数内で変数に付値しても，関数外には影響しない。関数内で変数の値を本当に変えてしまいたいときは，通常の付値でなくて，<<-（永続付値）を用いる。

2　人口ピラミッドを描くには

2005 年の日本の男女年齢各歳別人口データを使って人口ピラミッドを描くには次のコードを打てばいい[*3]。

```
x <- read.delim("http://phi.med.gunma-u.ac.jp/demography/jpop.txt")
library(pyramid)
pyramid(data.frame(Males=x$M2005/10000,Females=x$F2005/10000,Ages=x$Age),
 Laxis=0:3*50, Llab="男性",Rlab="女性",Clab="（歳）", Cstep=5, Cadj=-0.05,
 main="日本の人口ピラミッド，2005 年（単位：万人）")
```

日本のセンサスデータは，総務省統計局の web サイトで Excel 形式で提供されていて，そのままでは R で使えないので，複数のソースを組み合わせて整理し，タブ区切りテキスト形式で保存したものを自分の web サイトにアップロードした。データをパッケージ化して CRAN にアップロードすれば，もっと使いやすくなると思われる。

群馬県の 2005 年のセンサスデータだけは GunmaPop2005 というデータフレーム名で pyramid パッケージに入れてあるので，パッケージ呼び出し後であれば[*4]，pyramid(GunmaPop2005) とするだけで人口ピラミッドを描くことができる。

データさえ入っていれば，人口構造に関する指数の計算は式を打つだけでできるので，学生にも理解しやすい。2005 年の群馬県センサス人口から男女合計の老年化指数を求めるには次のコードを打つ。

```
attach(GunmaPop2005)
sum(Males[Ages>=65]+Females[Ages>=65])/sum(Males[Ages<15]+Females[Ages<15])*100
detach(GunmaPop2005)
```

すると [1] 142.7795 と表示される。日本全体の方は，

```
x <- read.delim("http://phi.med.gunma-u.ac.jp/demography/jpop.txt")
sum(x$M2005[66:86]+x$F2005[66:86])/sum(x$M2005[1:15]+x$F2005[1:15])*100
```

と打つと，[1] 146.5194 と表示される[*5]。

3　生命表の計算

東京大学で 2009 年度に死亡の分析として配布した資料（若干改訂した）ものを web に公開してある[*6]ので，詳細はそちらを参照されたい。

R 本体には生命表の計算をする関数は含まれていないが，年齢別死亡率のベクトルを引数として与えると生命表を返すような関数定義は 10 行程度で可能であり（ただし Greville の方法による補正を入れようと思うと 10 行では済まない），生命表の計算のしくみを学生に理解させるためには，自分で関数定義させる方がよい。Lx から Tx を求めるところで累積和（cumsum() という関数）を使う必要があるが，それ以外は難しくない。例えば以下の関数定義が可能である。

```
lifetable <- function(mx,class=1) {
 nc <- length(mx); qx <- numeric(nc); qx <- mx/(1+mx/2)
 dx <- numeric(nc); lx <- numeric(nc); Lx <- numeric(nc)
 lx[1] <- 100000
 for (i in 1:(nc-1)) {
  dx[i] <- lx[i]*qx[i]*class; lx[i+1] <- lx[i]-dx[i]; Lx[i] <- (lx[i]+lx[i+1])/2*class
 }
 Tx <- cumsum(Lx[nc:1])[nc:1]; ex <- Tx/lx
 data.frame(mx,qx,lx,Lx,Tx,ex)
}
```

Excel ではデータが入っているワークシートに生命表関数を入れる列を作り，式をコピーしていくことになり，計算式とデータが分離していないが，

[*3] http://phi.med.gunma-u.ac.jp/demography/makepyramid.html に各種応用例を解説してある。

[*4] ただし，version 1.2 の段階では，GunmaPop2005 の定義はコード部分ではなくドキュメント部分に入っているので，一度 example(pyramid) を実行しないと GunmaPop2005 の定義は実行されない。

[*5] 要旨集に書いた 154.1 はコトバンク等に載っていた推定値だが，確定数で計算すると 146.5 となる。

[*6] http://phi.med.gunma-u.ac.jp/demography/death.pdf

R では，いったん関数を定義してしまえば，どんなデータ（data という名前に付値されたベクトルだとする）についても，lifetable(data) と打つだけで生命表を得ることができる。一方，Keyfitz-Flieger の補整や Coale-Demeny のモデル生命表を使った補整も demogR パッケージの life.table() 関数には含まれているので，Excel では簡単にはできないような計算結果も得ることができる。

ヒトの人口学では，通常の年齢別死亡率 m_x（ある年に x 歳で死亡した人数 d_x をその年の x 歳年央人口で割った値）から q_x を $q_x = m_x/(1+m_x/2)$ として求めて生命表を計算するが[*7]，ゼロ歳のところは短期間での死亡率の変化が大きいので，最初の 1 ヶ月は 1 週間ずつ，次いで 2 ヶ月，3 ヶ月，6 ヶ月と刻んで，1 歳未満を 7 つの階級に分けて計算するのが普通である。また，年齢 5 歳階級で計算する生命表 (abridged life table) を作ることがあるが，その場合は死亡の線型性を仮定するのは無理なので別の補正を使う。よく用いられる Greville の方法は，

$$ {}_5q_x = \frac{{}_5m_x}{1/5 +_5 m_x \left[1/2 + 5/12 \left[{}_5m_x - \frac{\ln({}_5m_{x+5}) - \ln({}_5m_x)}{5}\right]\right]} $$

である (Ng and Gentleman, 1995)。この式は，5 歳階級での通常の年齢別死亡率の自然対数をとった値の傾き（$\frac{\ln({}_5m_{x+5}) - \ln({}_5m_x)}{5}$）が ${}_5m_x$ と等しければ，上の式と一致する[*8]。Greville の方法により補正を行う関数定義は次の通り[*9]。

```
lifetable <- function(mx,class=5) {
 nc <- length(mx)
 if (length(mx)!=nc) exit
 qx <- numeric(nc)
 for (i in 1:(nc-1)) {
  qx[i] <- mx[i] /
   (1/class+mx[i]*(1/2+class/12*(mx[i]-(log(mx[i+1])-log(mx[i]))/class))) / class
 }
 qx[nc] <- mx[nc] / (1/class+mx[nc]*(1/2+class/12*mx[nc])) / class
 dx <- numeric(nc); lx <- numeric(nc); Lx <- numeric(nc)
 lx[1] <- 100000
 for (i in 1:(nc-1)) {
  dx[i] <- lx[i]*qx[i]*class; lx[i+1] <- lx[i]-dx[i]; Lx[i] <- (lx[i]+lx[i+1])/2*class
 }
 Tx <- cumsum(Lx[nc:1])[nc:1]; ex <- Tx/lx
 data.frame(mx,qx,lx,Lx,Tx,ex)
}
```

ここで昭和 60 年の日本の 5 歳階級死亡データを使って Greville の方法を使った生命表を計算するには，web 上に公開している mortalityJ.R というコードを実行すると[*10]，S60ASMR という変数に年齢 5 歳階級別の mx が得られるので，続けて次のように打てばいい。

```
lifetable(S60ASMR,class=5,mode=2)
```

すると，次の結果が表示されるはずである。

[*7] この式の意味は次の通りである。年央人口で x 歳だった集団 N_x の実際の年齢は，x 歳以上 $x+1$ 歳未満である。もし死亡が一定速度で起こるなら，期首人口で x 歳だったのは $N_x + d_x/2$ 人となるはずである。この人たちが $x+1$ 歳になるまでに d_x 人死亡することになるので，死亡率 q_x は，

$$ q_x = \frac{d_x}{N_x + d_x/2} = \frac{d_x/N_x}{N_x/N_x + d_x/2/N_x} = m_x/(1+m_x/2) $$

となる。

[*8] Greville の式については，文献によって表記方法が若干異なるが，この，Ng and Gentleman (1995) による説明が一番わかりやすかった。オリジナルの Greville TNE (1943) Short methods of constructing abridged life tables. *Rec. Am. Inst. Actuar.*, 32: 29-43. が入手できないので確認できないのだが，舘 (1963) によれば，

$$ {}_nq(x) = \frac{{}_nm(x)}{\frac{1}{n} +_n m(x)\left[\frac{1}{2} + \frac{n}{12}\{{}_nm(x) - \ln c\}\right]} $$

で，c は ${}_nm(x)$ が Gompertz 法則に従うものとして ${}_nm(x) = Bc^x$ としたことにより，US の経験から $\ln c$ の値は 0.080 ～ 0.104 で，通常 0.09 が用いられるとされている。一方，和田 (2006) では，

$$ {}_nq_x = \frac{{}_nm_x}{\frac{1}{n} +_n m_x\left[\frac{1}{2} + \frac{n}{12}\{{}_nm_x - \log_e(\frac{{}_nmx+n}{{}_nm_x})^{\frac{1}{n}}\}\right]} $$

と書かれている。これは Ng and Gentleman とまったく同値である。生命表解析の専門書である，Namboodiri and Suchindran (1987) によれば，

$$ {}_nq_x = \frac{{}_nM_x}{(1/n) +_n M_x[(1/2) + (n/12)({}_nM_x - k)]} $$

かつ k は生命表によってわずかに異なるかもしれない定数で，0.09 に等しいとしても大きな誤差はないだろうとされている。数理人口学者 Keyfitz による “Applied Mathematical Demography” では，

$$ \int_0^n \mu(x+t)dt = n_nm_x + \frac{n^3}{12}{}_nm_x{}^2(\log_n m_x)\prime $$

が Greville の良く知られた結果であると書かれているが，n^3 の部分が他の式と違っていて，何を意味するのかよくわからないが，それを除けば，どの本も違いがあるのは 1 ヵ所で，そこをどの程度近似的な表現にしているかの違いなのではないかと思う。

[*9] source("http://phi.med.gunma-u.ac.jp/demography/lifetable.R") で，線型補正と Greville の方法を mode というオプションで切り替えられるようにした関数定義 lifetable() を読み込める。

[*10] source("http://phi.med.gunma-u.ac.jp/demography/mortalityJ.R") とすればいい。

```
                mx           qx        lx       Lx        Tx        ex
[0-4]   0.0014524735 0.0014455260 100000.00 498193.1 7602349.2 76.023492
[5-9]   0.0002099156 0.0002098010  99277.24 496125.8 7104156.1 71.558761
[10-14] 0.0001642103 0.0001641547  99173.09 495662.0 6608030.2 66.631280
[15-19] 0.0004690423 0.0004685105  99091.70 494878.2 6112368.3 61.683960
[20-24] 0.0005693208 0.0005685192  98859.57 493595.3 5617490.1 56.822927
[25-29] 0.0006039882 0.0006031087  98578.55 492149.6 5123894.8 51.977786
[30-34] 0.0007436492 0.0007423446  98281.28 490494.4 4631745.2 47.127440
[35-39] 0.0010362265 0.0010337771  97916.49 488317.2 4141250.8 42.293701
[40-44] 0.0017388068 0.0017318457  97410.37 484943.1 3652933.6 37.500459
[45-49] 0.0027567075 0.0027393406  96566.87 479527.7 3167990.5 32.806183
[50-54] 0.0045192235 0.0044715872  95244.22 470897.5 2688462.8 28.227043
[55-59] 0.0065107143 0.0064121767  93114.76 458110.4 2217565.3 23.815401
[60-64] 0.0094052904 0.0092050297  90129.42 440276.5 1759454.9 19.521427
[65-69] 0.0154376342 0.0149082601  85981.20 413883.1 1319178.3 15.342637
[70-74] 0.0269410609 0.0253626487  79572.05 372633.3  905295.2 11.377050
[75-79] 0.0486361813 0.0436233035  69481.26 309518.8  532661.9  7.666268
[80-84] 0.0862337753 0.0714027853  54326.25 223143.2  223143.2  4.107465
[85-]   0.1652433121 0.1124108423  34931.02      0.0       0.0  0.000000
```

他には，Gompertz 曲線を使って補正したり[*11]，Siler モデルを使って補正する方法もある（非線形最小二乗法でパラメータ推定すればよい。R での非線形最小二乗法は，`nls()` 関数や `optim()` 関数で実行できる）。なお，Siler モデルの式は，

$$h(t) = a_1 \exp(-b_1 t) + a_2 + a_3 \exp(b_3 t)$$

である。前出の公開している資料にはその方法も記載してあるが，大学生を対象とした講義では高度すぎるので，実際に説明するには至っていない。

4　まとめ

R による人口分析の講義・演習を Excel による場合と比較した際の利点と欠点は，次のようにまとめられる。

利点　数式を明示的に示す（とくにベクトル計算を式として扱える）点，モデルの当てはめやシミュレーションが容易である点，多くの対象について同じような計算を繰り返すことが容易である点

欠点　直感的に操作できるとは言い難い点，簡単な例から使い方を学ばねばならない点

従って，Excel よりも初学者にとっての障壁は大きいかもしれないが，今後，人口分析のためのパッケージが増えれば，R の利用はさらに容易になるであろう。また，分析例として使えるデータもパッケージに含まれているので，別にデータを用意しなくても操作法を提示することができ，教育目的に有用である。今後，権利関係をクリアし，日本の人口データをパッケージ化して CRAN に追加していくことができれば，より便利になるであろう。

5　補足

R の追加パッケージで人口分析関係のものは，本稿で触れた `Epi`，`demogR`，拙作 `pyramid` の他，既に CRAN に入っている `popbio` と，まだ CRAN に入っていない（2010 年 7 月に入る予定らしい）`demography`[*12]がある。教育目的には日本のデータでの使用例を含めたドキュメントが必要と思われるので作成中である。

[*11] `http://www.toukei.metro.tokyo.jp/seimei/2005/sm-gaiyou.htm` に示されている東京都の生命表は，高齢者の死亡が Gompertz 曲線で補正されている。

[*12] `http://robjhyndman.com/software/demography` に詳細な説明があるが，Rob J. Hyndman や John Maindonald が開発していて，UCB の John R. Wilmoth がやっている Human Mortality Database [`http://www.mortality.org/`，略して Berkeley Mortality Database ともいう。確か最初は 4 ヶ国だったと思うが，いつの間にか 37 の国または地域をカバーするまでに充実している] も入っているし，パッケージのサブタイトルにも書かれているように出生率と死亡率の将来予測をする関数が入っている。