情報処理第１回「データ入力」（A・B）

中澤　港

2004 年 10 月 4 日

データ入力のいろいろ

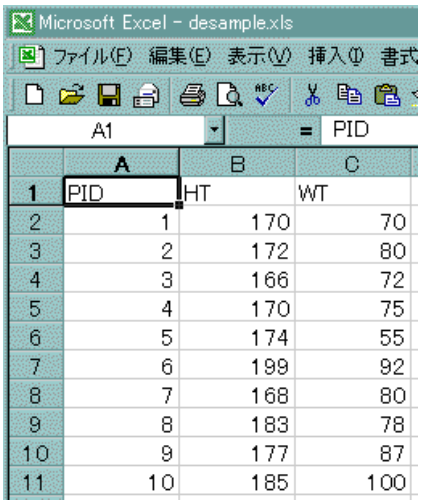
研究によって得られたデータをコンピュータを使って統計的に分析するためには，まず，コンピュータにデータを入力する必要がある。データの規模や利用するソフトウェアによって，どういう入力方法が適当か（正しく入力でき，かつ効率が良いか）は異なってくる。

さらに，入力以前に，コーディングとエディティング，それに，データクリーニングがきちんとできていないと，いくら正しく入力しても意味がない。コーディングとは，調査や実験によって得た生のデータを，どういう形の変数として保存するかを決めた一定の規則である。

ごく小さな規模のデータについて単純な分析だけ行う場合，電卓で計算してもよいし，分析する手続きの中で直接数値を入れてしまってもよい。例えば，60 kg, 66 kg, 75 kg という 3 人の平均体重を求めるには，Microsoft Excel では，１つのセルの中に`=AVERAGE(60,66,75)` とか`=(60+66+75)/3` と打てばいいし，R ならばプロンプトに対して `mean(c(60,66,75))` または `(60+66+75)/3` と打てばいい。

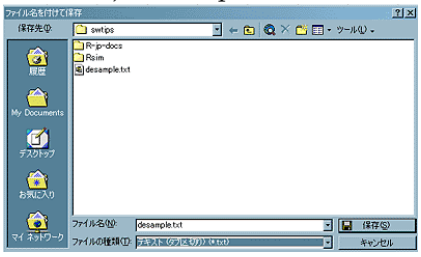
しかし実際にはもっとサイズの大きなデータについて，いろいろな分析を行う場合が多いので，データ入力と分析は別々に行うのが普通である。そのためには，同じ調査を繰り返しするとか，きわめて大きなデータであるとかでなければ，Microsoft Excel のような表計算ソフトで入力するのが手軽であろう。きわめて単純な例として，10 人の対象者についての身長と体重のデータが次の表のように得られているとする。

対象者 ID	身長 (cm)	体重 (kg)
1	170	70
2	172	80
3	166	72
4	170	75
5	174	55
6	199	92
7	168	80
8	183	78
9	177	87
10	185	100



まずこれを Microsoft Excel などの表計算ソフトに入力する。一番上の行には変数名を入れる。日本語対応 R なら漢字やカタカナ，ひらがなも使えるが，半角英数字（半角ピリオドも使える）にしておくのが無難である。この例の場合，対象者 ID を PID，身長を HT，体重を WT とするといい。入力が終わったら，一旦，そのソフトの標準の形式で保存しておく。入力完了した状態は，右の画面のようになる。

次に，この表をタブ区切りテキスト形式で保存する。Microsoft Excel の場合，メニューバーの「ファイル (F)」から「名前を付けて保存」を選び，現れるウィンドウの一番下の「ファイルの種類 (T)」のプルダウンメニューから「テキスト（タブ区切り）(*.txt)」を選ぶと，自動的にその上の行のファイル名の拡張子も xls から txt に変わるので，「保存 (S)」ボタンを押せば OK である（下のスクリーンショットを参照)。複数のシートを含むブックの保存をサポートした形式でないとかいう警告が表示されるが無視して「はい」を選んでよい。その直後に Excel を終了しようとすると，何も変更していないのに「保存しますか」と聞く警告ウィンドウが現れるが，既に保存してあるので「いいえ」と答えてよい（「はい」を選んでも同じ内容が上書きされるだけであり問題はない)。この例では，desample.txt ができる。



あとは R で読み込めばいい。この例のように，複数の変数を含む変数名付きのデータを読み込むときは，データフレームという構造に付値するのが普通である。保存済みのデータが `d:\desample.txt` だとすれば，R のプロンプトに対して，

```
dat <- read.delim("d:/desample.txt")
```

と打てば，データが dat というデータフレームに付値される。確認のためにデータを表示させたければ，ただ dat と打てばいいし，データ構造を見たければ，str(dat) とすればよい。３つの数値型（numeric，ただしこの場合はどれも整数の値しかないので整数型 integer になっている）変数として PID，HT，WT が読み込まれたことがわかる。変数の型には，この他に，カテゴリ変数に対応する要因型（factor），順序変数に対応する順序型（ordinal），真 (True) か偽 (False) かを示す論理型 (logical) などがあり，型によって適用可能な分析方法が違ってくる。数値型変数を無理やり順序扱いとか要因扱いすることは可能だが，逆は不可能な場合が多い。要因型変数に対して数値型変数にしか使えないような分析法を使いたいときは，ダミー変数化という方法を使うことができる。

読み込まれた変数に対して分析したいとき，例えばこの例の身長の平均と標準偏差を出したければ，

```
cat("mean=",mean(dat$HT),"sd=",sd(dat$HT),"\n")
```

とすればよい。一々 dat$と打つのが面倒ならば，attach(dat) とすれば，それ以降のセッション中，detach(dat) するまで，dat$を入力しなくても良くなる。例えば，このデータで身長と体重の相関係数を出して検定したいときは次のようにする。

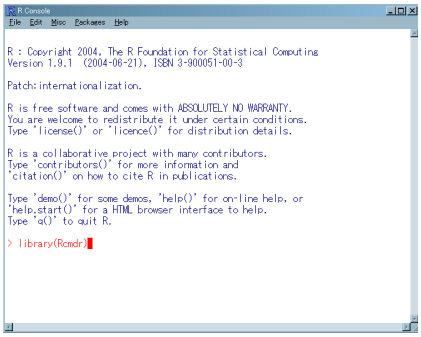
```
attach(dat)
cor.test(HT,WT)
detach(dat)
```

なお，R コマンダー（Rcmdr）を使って，次のようにメニュー形式でデータファイルを読み込むこともできる。

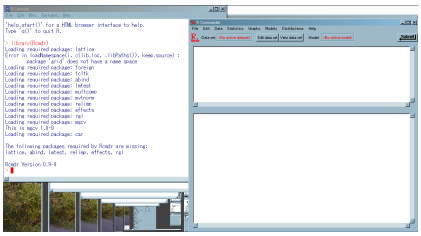
他のライブラリを使えるようにするときも基本的に同じだが，R Commander ライブラリを呼び出すには，

```
library(Rcmdr)
```

と入力する。エラーが１つでるが無視してよい。



他のライブラリとは異なり，R Commander ライブラリは，呼び出すだけで R Commander のウィンドウが起動する。



この状態から，先ほど保存した `d:\desample.txt` を読み込むためには，メニューバーの Data から Import Data の From Text File を開いて，Enter name for data set:の欄に適当な参照名をつけ（変数名として使える文字列なら何でもよいのだが，デフォルトでは Dataset となっている），Field Separator を White space から Tabs に変えて（Tabs の右にある○をクリックすればよい），OK ボタンをクリックすればよい。後は Rcmdr のメニューから選んでいくだけで，いろいろな分析ができる。

なお，データ入力は，入力ミスを防ぐために，2 人以上の人が同じデータを入力し，それを比較するプログラムを実行して誤りをチェックする方法がよいとされる。しかし，現実には 2 人の入力者を確保するのが困難なため，1 人で 2 回入力して 2 人で入力する代わりにするか，あるいは 1 人で入力してプリントアウトした結果を元データと見比べてチェックするといった方法が使われることも多い。

課題

`http://phi.med.gunma-u.ac.jp/medstat/p01.xls` をエクセルに読み込んでタブ区切りテキストファイル形式で出力し，含まれているすべての変数について，データの数と型を述べよ。

結果は学籍番号を含め，A4 の紙に印刷し，氏名を自筆して提出すること。結果の提出をもって出席確認とする。