医学情報処理演習第１回「データ入力」 2005年10月3日 中澤　港

## 統計処理ソフトの選択

医学情報処理として，この演習が目指すところは，実験，臨床，調査などで得たデータの統計解析ができるようになることである。現在では，統計処理はほとんどコンピュータを使って行われるので，まずは，適切なソフトウェアを選択する必要がある。しかし，ソフトウェアは無数にあって，どれを使うのが適当なのかわからない，という人も多いと思われる。

１つの方針としては，所属する研究室や身の回りで使っている人が多いソフトウェアを使うというのは合理的である。データを共用することもできるし，わからなくなったときに，すぐに誰かに尋ねることができる。この方法の欠点は，研究室を移ったときにそのメリットが失われることと，何か新しいことにチャレンジする場合に（いうまでもないが，論文を書くためには，研究のどこかが新しくなくてはいけない），自分でやり方を探索せねばならず，上述のメリットがあまりないことである[*1]。

では，とくにそういう制約がないとしたら，どういうソフトウェアがいいだろうか。下表は，国際的によく使われているソフトウェアを比較したものである。R をお勧めする理由は明白であろう。

| 比較項目 | SAS | SPSS | JMP | Excel | EpiInfo | R |
|---|---|---|---|---|---|---|
| メニュー操作 | ○ | ◎ | ◎ | △ | ◎ | ○ |
| プログラム実行 | ◎ | ○ | ○ | ○ | ○ | ◎ |
| 統計手法 | ◎ | ○ | ○ | × | ○ | ◎ |
| 作図能力 | △ | ○ | ○ | ○ | ○ | ◎ |
| 信頼性 | ◎ | ○ | ◎ | △ | ○ | ○ |
| 価格 | × | △ | △ | ○ | ◎（無料） | ◎（無料） |
| 動作 OS | ○ | △ | △ | △ | × | ◎ |
| 解説書 | ○ | ◎ | △ | ◎ | △ | ○ |

## データ入力

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，エディティングとコーディ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。エディティングとは，生データをエラーチェックなど精査して，回答そのものをチェックし，コーディングできるようにする過程であり，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を定めることである。厳密にやる場合は，コード表（調査票上の回答カテゴリーや分析機器からプリントアウトされた生の数値と入力すべきデータとの項目ごとの対応表）をつくり，データ形式も記載することが多い。コード表に基づいて調査票やプリントアウト上にコードを振ってから，初めてデータ入力ができることになる。

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，

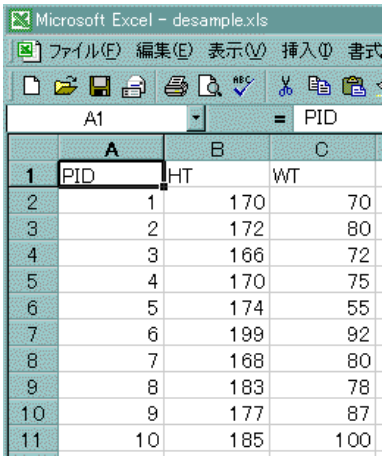
[*1] なお，Mac OS で動作する StatView というソフトは，簡便な操作性もあって臨床系で良く使われてきたが，SAS 社に買収された後，昨年末で開発・販売が停止されるに至り，今後サポートが期待できないので，これから統計解析を始める人は使わない方が無難であろう。

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。分析には R を使うとした場合，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[*2]。

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，R で使用するために，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。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。

## 欠損値について

ここで注意しなければならないのは，欠損値の取扱いである。一般に，統計処理をする対象のデータは，母集団から標本抽出したサンプルについてのものである。サンプルデータを統計解析して，母集団についての情報を得るためには，そのサンプルが正しく母集団を代表していることが何より大切である。質問紙調査の場合でも，実験研究の場合でも，欠損値（質問紙なら無回答，非該当，わからない，等，実験研究なら検出限界以

[*2] 大きなデータや繰り返し調査などの場合は，html でフォームを書き，httpd サーバソフトを動作させ，cgi を使ってデータ入力するとか，ACCESS などのデータベースソフトを使って入力フォームを設計して入力する方が間違いがないし効率も良い。

下，サンプル量不足，測定失敗等）をどのように扱うかによって，サンプルの代表性が歪められてしまうことがある。欠損が少なければあまり気にしなくていいが，たとえば，健診の際の食生活質問等で，「甘いものが好きですか」に対して無回答の人は，好きだけれどもそれが健康に悪いと判断されるだろうから答えたくない可能性があり，その人たちを分析から除くと，甘いもの好きの人の割合が，全体よりも少なめに偏った対象の分析になってしまう。なるべく欠損が少なくなるような努力をすべきだけれども，どうしても欠損のままに残ってしまった場合は，結果を解釈する際に注意する。

欠損値のコードは，通常，無回答 (NA) と非該当と不十分な回答が区別できる形でコーディングするが，ソフトウェアの上で欠損値を欠損値として認識させるためのコードは，分析に使うソフトウェアによって異なっているので（欠損値を表すコードの方を変更することも可能），それに合わせておくのも 1 つの方法である。デフォルトの欠損値記号は，R なら NA（R で読み込めるように，欠損値を NA にしてタブ区切りテキスト形式で出力したファイルはこれ），SAS なら.（半角ピリオド）である。Excel ではブランク（何も入力しない）にしておくと欠損値として扱われる，入力段階で欠損値をブランクにしておくと，「入力し忘れたのか欠損値なのかが区別できない」という問題を生じるので，入力段階では決まった記号を入力しておいた方が良い。その上で，もし簡単な分析まで Excel でするなら，すべての入力が完了してから，検索置換機能を使って（Excel なら「編集」の「置換」。「完全に同一なセルだけを検索する」にチェックを入れておく），欠損値記号をブランクに変換すれば用は足りる。

次に問題になるのが，欠損値を含むデータをどう扱うかである。結果を解釈する上で一番紛れのない方法は，「1 つでも無回答項目があったケースは分析対象から外す」ということである（もちろん，非該当は欠損値ではあるが外してはならない）。その場合，統計ソフトに渡す前の段階で，そのケースのデータ全体（Excel 上の 1 行）を削除してしまうのが簡単である（もちろん，元データは別名で保存しておいて，コピー上で行削除）。質問紙調査の場合，たとえば 100 人を調査対象としてサンプリングして，調査できた人がそのうち 80 人で，無回答項目があった人が 5 人いたとすると，回収率 (recovery rate) は 80%（80/100) となり，有効回収率 (effective recovery rate) が 75%（75/100）となる。調査の信頼性を示す上で，これらの情報を明記することは重要である。目安としては 80%程度は欲しい。

## R の基本操作

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[*3]。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

[*3] 演習室では，パスをつけなければすべての作業が 1 つの作業ディレクトリに対して行われるので，`dat <- read.delim("desample.txt")` とファイル名のみでいいはずである。

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。

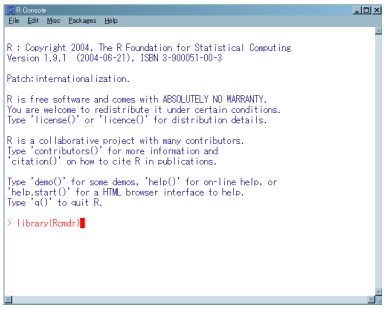
```
attach(dat)
cor.test(HT,WT)
detach(dat)
```

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。




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せよ。その後，含まれているすべての変数について，データの数と型を配布する紙に記入して提出せよ。結果の提出をもって出席確認とする。