科学研究におけるデータの取り扱いについて

東北大学大学院医学系研究科循環器内科学分野

文責　宮田　敏（PHS:6946）

2013/05/08

データを解析する際には、様々な解析手法が用いられます。しかし、どのような解析を行うとしても、その前に必ずやらなければならないこと、逆に絶対にやってはいけないことがあります。最低限やるべきことをルーチンワークとして行うだけで、データ解析のミスを大幅に減らすことができます。以下に、私が普段気を付けていることを纏めました。いささか長文ですが、御一読いただき参考にしていただければ幸いです。

1.　元データの取り扱い

i.　データの形は長方形。

データを入力する際は、第一行目に変数名を記入します。多くのソフトウエアは日本語入力に対応していますが、それでも全角文字は避けるほうが無難です。

第二行目以降にデータを記録していきますが、元データにはグラフ等を張り付けたりしません。また、第一列目にはデータの ID を記録します。そうすると、元データは以下のような長方形になるはずです。

systemID	hospitalID	sex	age	height	bodyweight
4	1185645	1	64	173	75.4
11	3329388	1	69	164	72
12	4022624	1	78	155.2	47.2
14	4402536	1	83	159.1	60
22	4862866	2	73	147.6	40.5

ii.　元データは絶対に改変しない。

データを解析する際、変数を変換したり新しい変数を定義したりする必要が出てくることがあります。このとき元データを改変して、変換した変数を上書きしたり変数を新たに保存したりしてはいけません。データを改変したときは、必ず新しいファイル名で保存しなおします。元のデータを改変した場合、解析を進めるうちに元データが何であったのか分からなくなることがあります。元データがわからなくなれば、意図せざるデータのねつ造まであと一歩です。

iii. 患者さんの**個人情報は記載しない**。

個人情報保護の重要性は改めて述べるまでもありませんが、残念ながらいまだに患者さんの名前やカルテ番号など個人に直結するデータを記録したままでデータをやり取りする例が見受けられます。患者さんの個人情報は、データ解析の立場から見れば何の意味もありませんが、万が一外部に流出する、あるいは記録媒体を紛失するなどすれば、研究の中断では済まない問題に発展します。

解析データの**個人情報は削除**する、を徹底する必要があります。

iv. 解析の過程の**詳細なメモ**を残す。

様々な実験の結果を論文にまとめる際、「実験ノート」に詳細を記録することは常識ですがデータ解析でも同じです。解析の過程を記録するのには、次のような意味があります。

- **研究の再現可能性**を確保するため。
  科学的研究においては、第三者の事後的な検証に耐えられるよう研究の再現が可能でなければなりません。元データと記録メモさえあれば、それ以外に知識のない人でも解析が再現できるような詳細なメモを残す必要があります。
- **備忘録**。自分自身、何をしているのか分からなくなることを防ぐため。
  一日数時間の解析でデータ解析が終わることは、まずありません。一か月、二か月と時間をかけて解析を進めた場合、最初のころに自分が何をしていたのか分からなくなることがあります。データ解析の世界には、**「三日後の自分は遠い親戚、一週間後の自分は赤の他人」**という言葉があります。赤の他人が見ても、何をしているのか分かるようなメモを心がけます。

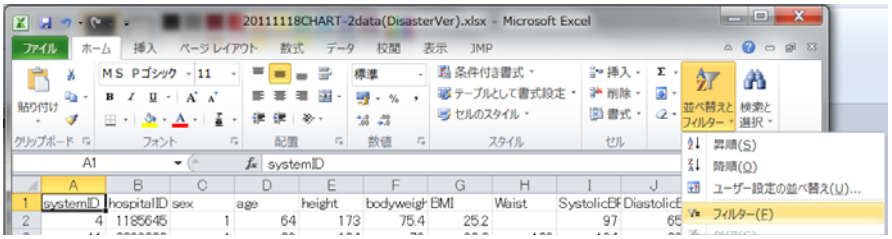
2. データ入手時にまずすべきこと

i. データ全体の確認

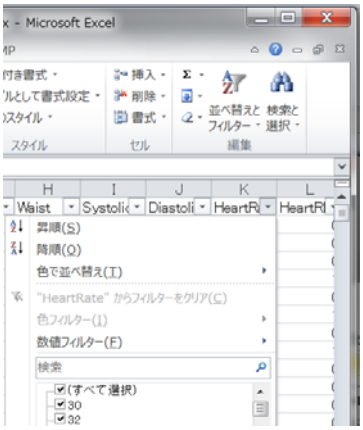
最初に得られた生データには、往々にして記録ミスや不適切な入力が存在するものです。そういった誤りは適切に修正、削除する必要がありますが、その確認作業を体系的に行うことでミスを減らし時間を節約することができます。

以下の手順は、私が普段行っているものですが参考にして頂ければと思います。

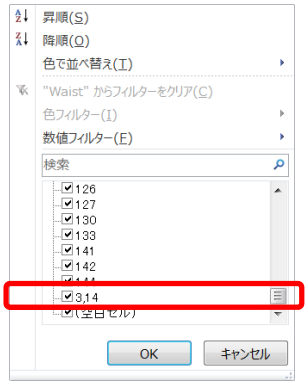
- Excel でデータファイルを開き、「並べ替えとフィルター」→「フィルター」を押してフィルターをオンにする（列見出しに矢印が現れる）



- 列見出しの矢印をクリックして、各列に適用するフィルターが見えるようにする。



- 以下の要領でデータの誤りや異常値の有無を確認する
  - データの範囲：本来、正値しか入らないはずの変数に、負値が入っている等。
  - データが想定の範囲を逸脱する。（小数点の桁間違いで、例えば体重 35.0 kgが 3.5 kgと入力されるなど。）
  - 全角文字と半角文字の混在。（“I” と “I” の区別など難しい。前が半角、後ろが全角）
  - 質的変数の数字表記：例えば「性別」が、男性→1, 女性→2 で記入されるような場合があるが、間違いの元なので、男性→M, 女性→F のように書き直す。
  - 異常な値の検出:例えば小数点とカンマの打ち間違いで、“3.14” が “3,14” となっている場合など。そういった異常値はフィルターの下の方に出る。



  - 欠測値の数：データに欠測がある場合、フィルターに「空白セル」と表示される。「空白セル」を選択してフィルターをかければ、欠測値の個数を調べられる。欠測の数が想定より大きかった場合、入力したデータが認識されていない、などの可能性が考えられる。

ii. 以上の確認作業を、すべての変数について行う。どのデータに対して、いかなる修正、削除を行ったか、すべて解析メモに記録する。修正後のデータは新しいファイル名で保存し、これを解析ファイルとする。（元データには手を付けない）

3. 解析メモの作成要領（テンプレート）

前述の 1-iv で書いた通り、データ解析の過程を記録したメモは解析ミスを減らすため、また後からミスを発見するためにも大変に有用です。解析の途中でいちいち記録を残すことは一見面倒のように思えますが、結局は解析時間の短縮にもつながります。以下、普段私がメモを作るときに記録している項目をまとめてみました。

i. **データ入手の経緯。**データをメールで受け取った場合は、以下のような感じ。自分の実験データの場合も、データ入手年月日と実験内容、実験実施者等を記録。

【データ】　　2012/04/04 付で、**先生よりいただいた以下のファイル。

- SR_STUDY_1_2_ALL_VIEW.csv　　元ファイル
- 計 10843 症例　111 変数

ii. **データの修正、ファイル操作の記録。**前ページの要領でデータの誤りを発見した場合の、データの修正、削除の記録。このような操作は、できれば手動ではなく何らかのプログラムを組んで行いたいが、手動で行う際はなおのことファイル操作の一つ一つを記録する。（後から第三者がみて、同じ修正ができるように）
例えば、以下のような感じで記録する。

【ファイル操作】

1. Excel で SR_STUDY_1_2_ALL_VIEW.csv を開く。
2. “null” を “NA” に置換。
3. “TG” に年号が入っているセル（計 13 個）がある。すべて NA とする。
4. 「テキスト（タブ区切り）」形式で保存。SR_STUDY_1_2_ALL_VIEW.txt

iii. **解析内容の記録。**行った解析の内容を記録し、解析結果、作成したグラフ等を記録する。例えば SPSS で解析した結果をそのまま SPSS に保存したり、Excel で描いたグラフを Excel のワークシートに張り付けて保存したりする例がありますが、お勧めしません。それぞれ個別のファイルに保存し、ファイル名を記録することをお勧めします。ファイル名はどのようにつけてもかまいませんが、わたしは、“プロジェクト名” + “作業内容” + “日付” をファイル名にしています。
例：

- CHART2_StageCD_coxph_130430.txt
- CHART2_StageCD_性差_CrudeAllCauseDeath130226.png

4. 迷った時は相談を

解析の内容に**迷った時**、**自信がないとき**、解析結果の解釈に**困ったとき**等は、いつでも**ご相談ください**。それもなるべく早く、できれば解析を始める前にご相談ください。

Email: miyata@cardio.med.tohoku.ac.jp
PHS: 6946