【セミナー】

社会調査データの入力とチェックの方法

小林　久高・雨森　聡・山本　圭三
KOBAYASHI Hisataka, AMENOMORI Satoshi,YAMAMOTO Keizo

1　はじめに

計量的な調査研究においては、調査票などを用いて集められた情報をコンピュータで分析できるデータにする作業が不可欠である。分析に用いるデータは調査票をきちんと反映したものでなければならないので、この作業は丁寧かつ正確におこなわれる必要がある。綿密な計画の下で得られたデータであっても、この段階でミスを犯してしまっては、後の分析は信頼できるものではなくなってしまう。

この段階でのミスを最小限にするために、データ入力およびデータチェックの過程には、いくつかの段階が設けられるのが普通である（図 1）。すなわち、(1) 調査票を作成した後に、(2) 問いごとに変数名をつけ、(3) コードブックを作成したうえで、(4) データを入力する。そして、(5) 入力されたデータをチェックし、問題があった場合には再入力をおこなう。(6) そして、最後にロジカルエラーのチェックをする。こうした作業を経て、ミスの少ない分析用のデータが完成するのである。

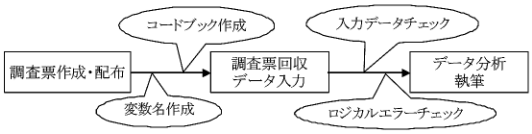


図 1　調査票作成からデータ分析までの過程

現在では、こうした一連の作業はすべてパソコンを使っておこなわれる。ところが、一般的な調査法のテキストにおいては、実際にパソコンを使ったデータ入力とデータチェックの実践的な方法まで記述したものはあまり見られない。そこで本稿では、調査票作成以降の過程について、具体的な方法を解説していくことにする。

2　作成されるデータの 2 つの形式

具体的な説明に入る前に、われわれが作成しようとしているデータの形式について説明しておこう。計量的な調査で得られるデータには、一般的に 2 つの形式がある。

1 つはデータをテキスト形式で入力する方法である（図 2）。

図 2

テキスト形式のデータは汎用性がある（特殊なソフトを用いなくても入力でき、多くの統計ソフトに利用できる）ため、用いられることが多い。だが、自由記述など文字型のデータの処理が面倒であったり、データのチェックをする段階になると少々複雑な作業を要する、といった欠点もある。

もう 1 つは、エクセルを用いてデータを入力する方法である（図 3）。

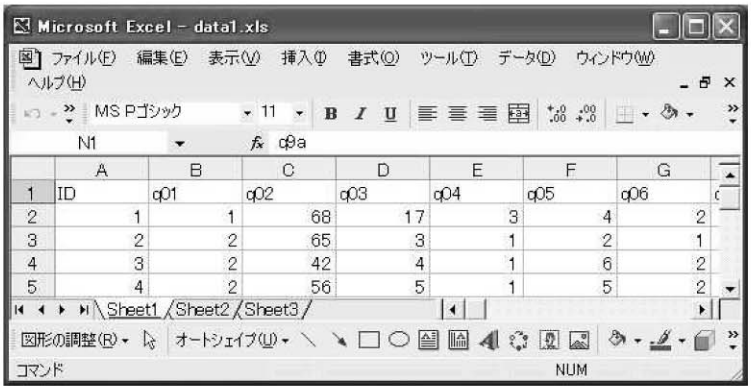


図 3

【問 1】あなたの性別を教えてください。q01

1.男　　2.女

【問 2】あなたの年齢と入学年度を教えてください。（学籍番号の上から 3・4 ケタ目が入学年度になっています。例 1204△△△△→2004 年度入学）

2006 年 4 月 1 日現在（ q02x1 ）歳　　西暦（ q02x2 ）年度入学

【問 3】あなたの現在の居住形態は次のうちのどれですか。1 つに○をしてください。q03

1.自宅に住んでいる　2.以前は下宿していたが、今は自宅に住んでいる　3.下宿（寮含む）に住んでいる

【問 4】あなたのご両親の年齢（2006 年 4 月 1 日現在）を教えてください。（離別・死別の場合は、「該当者はいない」に○をつけてください。）

父（ q04x1 ）歳／該当者はいない　母（ q04x2 ）歳／該当者はいない

図 4

エクセルで入力する場合、文字型のデータも処理しやすいし、データチェックの作業はテキスト形式のデータに比べて非常に簡単である。しかし、作業を進めるにあたってはエクセルに関する一定の知識が必要となる。このため、エクセルにある程度慣れていない者にとってはつらい作業になってしまう。

それぞれにメリットとデメリットがあるため、どちらが優れているとは一概に言えない。そこで、以下ではそれぞれの方法でデータを作成する作業を解説していくことにする。

3　調査票とコードブック、入力

3.1 調査票を作成したあと

われわれの出発点は調査票である。調査票作成後、まず行う作業は、調査票の各問に変数名をつけることである。この作業は、どのような形式でデータを入力にするかにかかわらず必要になる。

各問につけた変数名は、図 4 のような形で調査票に記入しておくと便利である。すなわち、【問 1】に q01、【問 2】の最初のカッコに q02x1、後のカッコに q02x2 という具合に変数名を記入するのである。調査票がワードなどで作られている場合、

「変数名付き調査票」という形で保存しておくといいだろう。

3.2 コードブック

各問に変数名をつけた後､次にすべき作業は｢コードブック」をつくることである。コードブックとは、端的に言えば回答済みの調査票の情報をデータとして入力していく際の規則を記したものである。入力する方法によって、コードブック必要な情報は異なる。以下で、それぞれの場合のコードブックを紹介しよう。

(1) テキストで入力する場合のコードブック

図 5 は、図 4 の調査票で得られるデータをテキスト形式で入力する場合のコードブックである。そこには、｢問番号」、｢調査項目」、｢変数名」といった情報とともに、｢無回答」であったり｢指定外」の回答があったり、回答者には該当しない「非該当」であったりした場合にはどのような値を入力すべきかが記されている。

テキストで入力する場合、最も重要になるのは「桁数」、「開始カラム」、「終了カラム」の情報である。図 2 に示したとおり、テキスト形式のデータは数値がずらっと並ぶ形になる。一目見ただけでは何のことかわからないが､数値はすべて「ID」の値→「問 1」の値→「問 2-1」の値、といった具合に規則正しく並べられる。したがって、並べられている値を各変数のデータとして認識するために重要になるのは、値がおかれている位置を表す「桁（カラム)」である。すなわち、何桁目にはどの変数の値を入れるのか、という規則をあらかじめ明記しておく必要があるのである。

回答してもらった調査票を回収した後、最初に通し番号を打つ。その後、このコードブックにしたがってデータ入力がなされることになる。1 行目にはケース 1 のデータ、2 行目には番号 2 のケースのデータと、桁数に注意しながらどんどんデータを入力していく（場合によっては、1 ケースのデータを複数の行に入力することもある)。

	A	B	C	D	E	F	G	H	I	J	
1	問番号	調査項目		変数名	桁数	開始カラム	終了カラム	無回答	指定外	非該当	
2		番号		ID	3	1	3	-	-	-	
3	1	性別		q01	1	4	4	9	7	-	
4	2	年齢		q02x1	2	5	6	99	97	-	
5		入学年度		q02x2	4	7	10	9999	9997	-	
6	3	居住形態		q03	1	11	11	9	7	-	
7	4	父年齢		q04x1	2	12	13	99	97	98	「該当者はいない」に
8		母年齢		q04x2	2	14	15	99	97	98	「該当者はいない」に

図 5

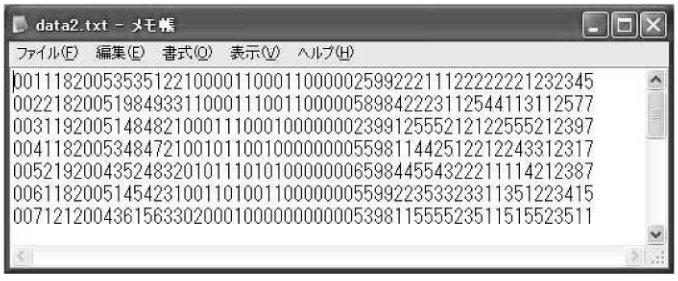


図 6

図 6 は、図 5 のコードブックにしたがって実際に入力したデータを示している（1 ケース 1 行）。図 6 のデータからは、たとえば、番号 001 のケースは、男性（4 桁目）で 18 歳（5〜6 桁目）、…父年齢は 53 歳（12〜13 桁目）、母年齢は 51 歳（14〜15 桁目）である、といったことが読み取れる。

（2）エクセルで入力する場合のコードブック

図 7 は、図 4 の調査票で得られるデータをエクセルで入力する場合のコードブックである。基本的な情報はテキスト形式の場合と同じであるが、大きく異なるのは、「エクセル列 1」、「エクセル列 2」の部分である。

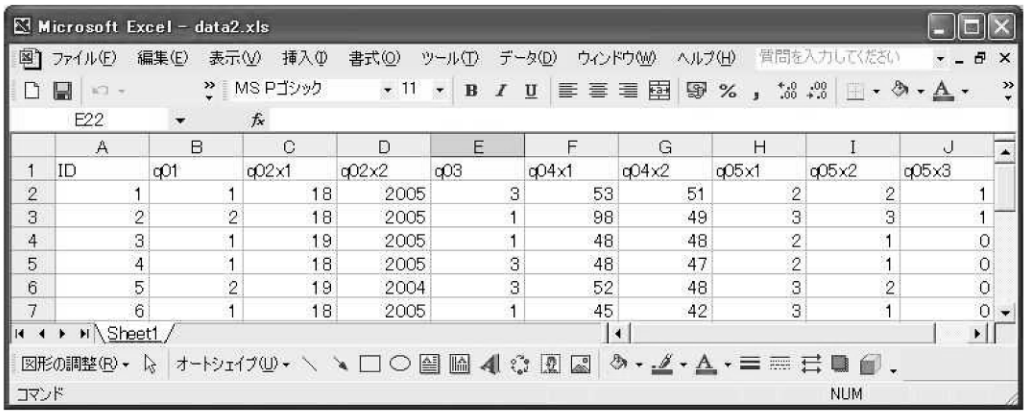
エクセルで入力する場合、図 3 のように値は変数ごとに別々の列に入力されることになる。このため、エクセルの場合は値を「何桁目に入力するか」ではなく「どの列に入力するか」が重要になる。エクセルの列番号はアルファベットの組み合わせで表されるため、図 7 のように「エクセル列 1」、「エクセル列 2」として明記しておくとよい。

実際に入力する際には、まずワークシートの 1 行目に変数名を入力し、列と変数の対応関係をよりわかりやすくしておく。その後、コードブックにしたがって 2 行目に番号 1 のケースのデータ、3 行目に番号 2 のケースのデータといった具合にデータを入力していけばよい。

図 8 は、図 7 のコードブックにしたがってデータを入力したワークシートの例である。ワークシートからは、たとえば番号 4 のケースは、女性（列 B）で 18 歳（列 C）、父年齢は 48 歳（列 F）、母年齢は 47 歳（列 G）であることなどが読み取れる。

	A	B	C	D	E	F	G	H	I	J	
1	問番号	調査項目		変数名	桁数	無回答	指定外	非該当	エクセル列1	エクセル列2	
2		番号		ID	3	−	−	−		A	
3	1	性別		q01	1	9	7	−		B	
4	2	年齢		q02x1	2	99	97	−		C	
5		入学年度		q02x2	4	9999	9997	−		D	
6	3	居住形態		q03	1	9	7	−		E	
7	4	父年齢		q04x1	2	99	97	98		F	「該当者はいない」に
8		母年齢		q04x2	2	99	97	98		G	「該当者はいない」に

図 7



	A	B	C	D	E	F	G	H	I	J
1	ID	q01	q02x1	q02x2	q03	q04x1	q04x2	q05x1	q05x2	q05x3
2	1	1	18	2005	3	53	51	2	2	1
3	2	2	18	2005	1	98	49	3	3	1
4	3	1	19	2005	1	48	48	2	1	0
5	4	1	18	2005	3	48	47	2	1	0
6	5	2	19	2004	3	52	48	3	2	0
7	6	1	18	2005	1	45	42	3	1	0

図 8

4　入力データのチェック

データ入力に関するミスで最も起こりやすいのは、入力時のミス（たとえば、数字の読み違い、キーの打ち間違いなど）である。これらを回避する最も単純な方法は、データを 2 度以上入力し、入力されたデータが同一であるかどうか照合することである。以下で、その方法を紹介しよう。

4.1 入力ミスのチェック

（1）テキストで入力する場合のチェック方法

テキスト形式のデータの場合、(1) データを 2 回入力して別々のファイルに保存する、(2) 2 つのファイルを照合し、違いのある部分を再入力する、という流れでチェックがおこなわれる。

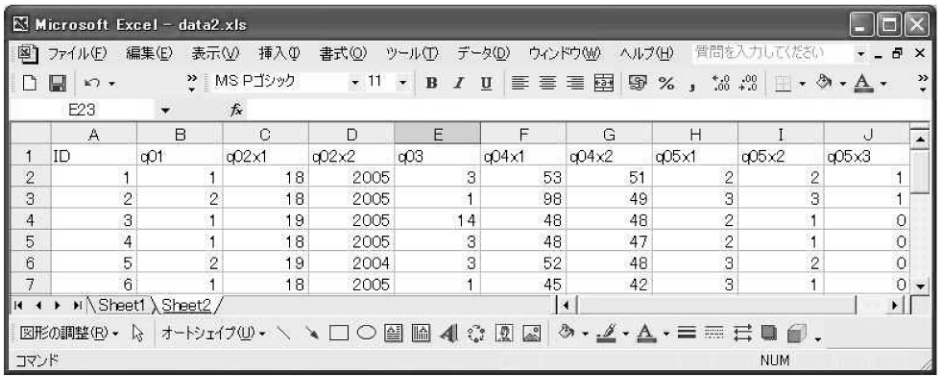
データの照合はプログラムを用いておこなうことが多いが、紙幅の関係上ここでは説明を省略する。データの照合と再入力について、筆者の 1 人が作成した C のプログラムに関する文献（小林 1994）があるので、詳しくはそちらを参照していただきたい。

（2）エクセルで入力する場合のチェック方法

エクセル形式の場合も、基本的な手順はテキスト形式の場合と同じである。まず、先ほど述べたように全回答者分のデータをエクセルの 1 枚のシートに入力していく（Sheet1=1 度目の入力）。次に、今度は別のシートに全回答者分のデータを入力していく（Sheet2=2 度目の入力）。2 度の入力が終われば、エクセルの表示は図 9 のようになる。

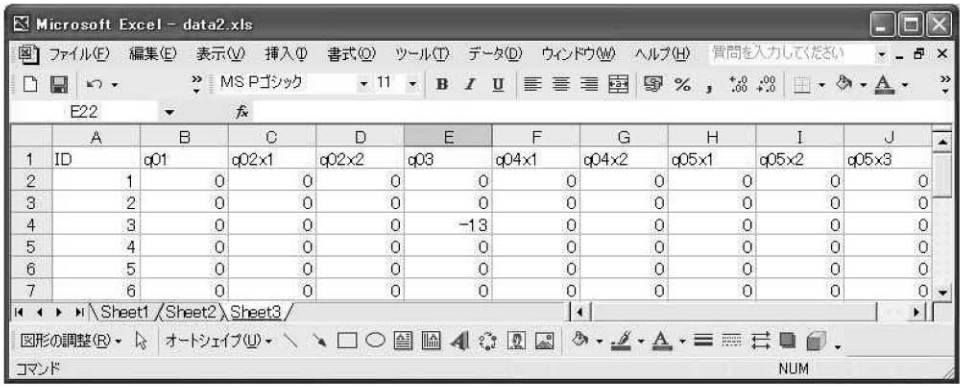
2 度の入力が済んだら、別のシート（Sheet3）において、入力したデータの照合をする。照合するには、Sheet1 と Sheet2 それぞれの同じセル番地についての引き算を Sheet3 ですればよい。たとえば、セル B2 の照合には、Sheet3 の B2 に「=Sheet1!B2－Sheet2!B2」という式を入力すればよい。この式をコピーし、シート全体に渡ってそれを貼り付ければ、データ全体の照合ができる。

引き算の結果が「0」の場合はミスなく入力されていることを、「0」ではない場合はどちらかのシートに入力ミスがあることを意味する。図 10（Sheet3）は、図 8（Sheet1）と図 9（Sheet2）の引き算の結果を示したものである。図 10 より、E4 のセルで入力ミスが起きていることがわかる。メニューの「書式→条件付き書式」を用いて Sheet3 全体に「0」でないセルに色をつけるようにすると、よりわかりやすくなるだろう。



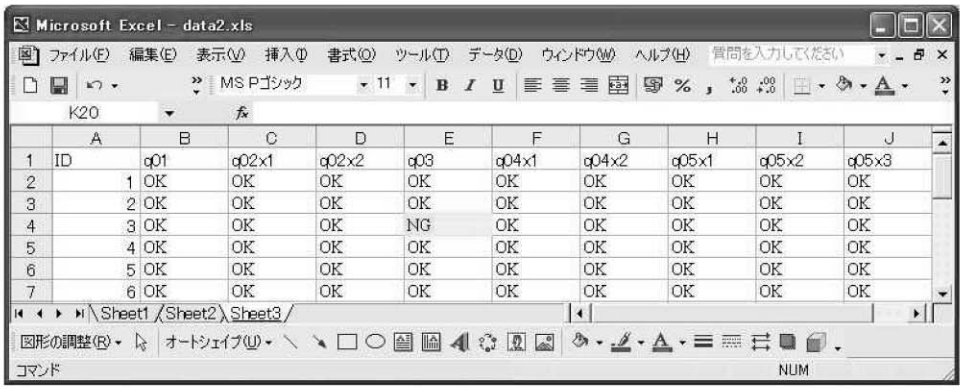
	A	B	C	D	E	F	G	H	I	J
1	ID	q01	q02x1	q02x2	q03	q04x1	q04x2	q05x1	q05x2	q05x3
2	1	1	18	2005	3	53	51	2	2	1
3	2	2	18	2005	1	98	49	3	3	1
4	3	1	19	2005	14	48	48	2	1	0
5	4	1	18	2005	3	48	47	2	1	0
6	5	2	19	2004	3	52	48	3	2	0
7	6	1	18	2005	1	45	42	3	1	0

図 9　2 回目の入力（Sheet2）



	A	B	C	D	E	F	G	H	I	J
1	ID	q01	q02x1	q02x2	q03	q04x1	q04x2	q05x1	q05x2	q05x3
2	1	0	0	0	0	0	0	0	0	0
3	2	0	0	0	0	0	0	0	0	0
4	3	0	0	0	-13	0	0	0	0	0
5	4	0	0	0	0	0	0	0	0	0
6	5	0	0	0	0	0	0	0	0	0
7	6	0	0	0	0	0	0	0	0	0

図 10　入力ミスのチェック例 1（E4 に間違いあり）



	A	B	C	D	E	F	G	H	I	J
1	ID	q01	q02x1	q02x2	q03	q04x1	q04x2	q05x1	q05x2	q05x3
2	1	OK	OK	OK	OK	OK	OK	OK	OK	OK
3	2	OK	OK	OK	OK	OK	OK	OK	OK	OK
4	3	OK	OK	OK	NG	OK	OK	OK	OK	OK
5	4	OK	OK	OK	OK	OK	OK	OK	OK	OK
6	5	OK	OK	OK	OK	OK	OK	OK	OK	OK
7	6	OK	OK	OK	OK	OK	OK	OK	OK	OK

図 11　入力ミスのチェック例 2（E4 に間違いあり）

文字データがセルに入力されている場合、シート間の引き算ではデータの照合ができない。このような場合は IF 関数（例えば、C4 の照合なら「= IF (Sheet1!C4=Sheet2!C4,"OK","NG")」と入力）を用いればよい（図 11）。

以上のようなデータのエラーチェックをおこなえば、入力ミスを少なくすることができる。ちなみに、2 度入力する際には、1 人で 2 度入力するよりも、2 人 1 組で 1 度ずつ入力するようにすればより効果的である。

4.2 ロジカルエラーのチェック

さて、回収された調査票には時にロジカルエラーが含まれている。ロジカルエラーというのは、たとえば、20 歳なのに「自分は 40 年その町に住んでいる」と答えていたり、40 歳なのに実母が 20 歳と答えていたりすることをいう。また、調査票には時にあるカテゴリーの人にだけ聞きたい質問が含まれる。そういうときには非該当者に対して「次は問○へお進みください」と指定がなされ、非該当質問をパスするような誘導がなされる。しかし、誘導は回答者に無視されることもある。このような場合も、ロジカルエラーと考えていい。

ロジカルエラーが生じているときどう対応すべきかは一概には言えない。問題によって対応方法は異なってくる。重要なのは、ロジカルエラーに対してどのような方法で対処したかをきちんと記録しておくことである。こうすることによって、データの再検討が可能になるからである。

以下では、ロジカルエラーのチェックの具体的な作業について説明する。まず、エクセルを用いた方法を解説しよう。

（1）エクセルで入力する場合のチェック方法

①データファイルをコピーする。

3 で述べた 2 枚のシートを照合して修正した「入力ミスのチェック済みデータファイル」は貴重品であり、これは必ず保管しておく必要がある。これをコピーしたロジカルエラーチェック用のファイルをまず作成する。

②オートフィルタを設定する。

ロジカルエラーチェック用のファイルを開き、メニューから「データ→フィルタ→オートフィルタ」を選ぶ。そうすると、図 12 のように▼印が 1 行目に表示される。

③フィルタをさまざまにかけてロジカルエラー発見し修正する。

この▼をクリックすると、さまざまなフィルタがかけられるので、それをもとにデータの修正を行う。たとえば q01 が 1 のものは q03 は非該当だとすると、フィルタを q01 にかけて「1」のものだけ表示するようにして、q03 がどうなっているかを確認し、適切に修正するのである。

④エクセルの計算式でデータをチェックする。

エクセルではさまざまな計算式が立てられる。シートに新たな列を作成し、ここにロジカルエラーを発見できる式を記入する。そして、この式をもとにロジカルエラーのチェックと修正を行う。もちろんこのときオートフィルタも併用することになる。

（2）テキスト形式⇔エクセル形式の変換方法

テキスト形式データのロジカルエラーチェックは、入力ミスのチェックと同様、プログラムを用いることになる。ところが、単純なデータの照合ではないのでプログラムも少々複雑になり、エクセルを用いるよりもはるかに厄介な作業になる。それゆえ、テキスト形式のデータであっても、エクセル形式のデータに変換し、上述の方法を用いてチェックするのが現実的である。テキスト形式のデータをエクセル形式に変換する方法と、一度エクセル形式に変換したデータをテキスト形式にもどす方法を解説しておこう。

	A	B	C	D	E	F	G	H	I	J
1	ID	q01	q02x1	q02x2	q03	q04x1	q04x2	q05x1	q05x2	q05x3
2	1	1	18	2005	3	53	51	2	2	1
3	2	2	18	2005	1	98	49	3	3	1
4	3	1	19	2005	1	48	48	2	1	0
5	4	1	18	2005	3	48	47	2	1	0
6	5	2	19	2004	3	52	48	3	2	0
7	6	1	18	2005	1	45	42	3	1	0
8	7	1	21	2004	3	61	56	3	3	0
9	8	2	18	2005	1	43	41	2	1	0

図 12　オートフィルタの例

①まず、変数と変数の間をカンマで区切る。

3.2 で述べたように、エクセル形式ではデータが変数ごとに別々の列に収まるようになっている。このため、まずすべてのケースについて 1 変数ごとに半角のカンマ（,）を入れ、変数ごとの区切りを設けなければならない（図 13）。

図 13　変数の間にカンマを入れたデータ

②カンマを入れたデータを、新しくカンマ区切り形式（csv）のファイルとして保存する

カンマを入れたデータを保存する際に、「名前を付けて保存」を選択し拡張子を「.csv」と指定すれば、データはカンマ区切り形式（csv）のファイルとして保存される。カンマ区切り形式のデータは多くの表計算ソフトに対応する形なので、エクセルが入っているパソコンであれば、カンマ区切り形式のファイルはエクセルで開くことができる。

以上のようにすれば、テキスト形式のデータがエクセル形式のデータに変換される。ちなみに、カンマ区切り形式のファイルは、そのままエクセルから開いてもファイル形式はカンマ区切りのままになる。エクセルのワークシートとして保存したい場合は、エクセル上で「名前を付けて保存」する際に拡張子を「.xls」と指定すればよい。

次に、いったんエクセル形式に変換したデータをテキスト形式に戻す方法について説明しよう。

①変数の桁表示を変更する

テキスト形式のデータがエクセル形式に変換されたとき、値のうち不必要な桁は削除されてしまう（たとえば、「001」は「1」、「018」は「18」となる）。このため、いったんエクセル形式にしたデータをそのままテキスト形式にもどすと、ケースによっては桁のずれが生じてしまう。

こうした事態を回避するためには、次の処理をおこなえばよい。桁を変更したい列を選択し、メニューから「書式→セル」を選ぶ。「セルの書式設定のダイアログ」が出てくるので、「表示形式」タブをクリックし、「分類」の欄にある「ユーザー定義」をクリックする（図 14）。

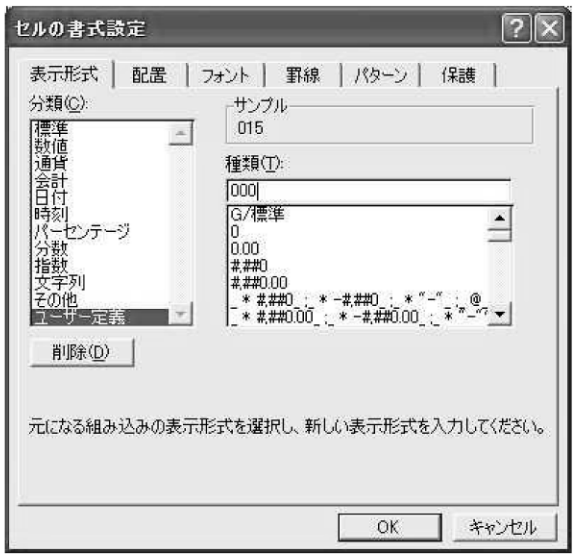


図 14

そして、ダイアログ右の「種類」のボックスに必要な桁数の分だけ「0」を直接入力する（たとえば、3 桁入力の変数であれば「000」と入力する）。このようにしておけば、1 は「001」、18 は「018」といった具合に表示されるようになるので、テキスト形式にもどしても桁のずれが生じることはな

くなる。

②エクセル上でデータをコピーし、メモ帳などに貼り付ける。

エクセル上でデータをすべてコピーしてメモ帳などに貼り付けると、変数と変数の間にタブが入った形（タブ区切り形式）で貼り付けられる（図15）。「置換」機能などを用いてこのタブを消去すれば、データは元のテキスト形式にもどる。

図 15　変数の間にタブが入ったデータ

5　SPSS データファイルの作成

ロジカルエラーが修正されたらファイルを保存し、そのファイルから SPSS で分析できるファイル（SPSS データファイル）を作成することになる。そのための手続きは次の通りである。

①SPSS を立ち上げる。
②シンタックスエディタを開く。

SPSS を起動させた後に、メニューから「ファイル→新規作成→シンタックス」と選択すれば、新規のシンタックスエディタが開く（図 16）。

図 16

データの形式によって、データの読み込みに必要なシンタックスは異なる。以下で、それぞれの場合のシンタックスを説明しよう。

5.1 テキスト形式のデータを読み込む場合

テキスト形式のデータの場合、次の内容をシンタックスエディタに記述することになる。

```
data list file='ファイルの場所¥ファイル名.txt'
    records=2
/1      変数 A   1
        変数 B   2-4
        変数 C   5
/2      変数 D   1-4.
```

「file」は、読み込むファイルのある場所とファイル名を指定するサブコマンドである。「records」は、1 ケースのデータが何行にわたっているかを示すものである。

「/1」と「/2」は、変数定義のサブコマンドである。先にも述べたように、テキスト形式では、数値が変数の区別なく並ぶかたちになっている。このため、どの変数が何桁目から何桁目までなのかを、変数ごとに指定しなければならない。変数の範囲を指定し、変数名を付ける作業をおこなうのが、「/1」と「/2」のサブコマンドなのである。

そして、「/1」は変数の定義の対象となっているデータの行を指定している。「/1」の隣にあるのが定義する変数の名前であり、後ろの「1」はその変数の値とする範囲を桁番号で指定するものである。したがって、「/1　変数 A　1」とは「1 行目の 1 桁目を『変数 A』と定義する」という意味である。変数の範囲が 1 桁でない場合は、「2-4」のように間をハイフン (-) でつないで範囲の指定をすればよい。データの行が複数に渡っている場合は、例に示した「/2」のように、適宜スラッシュと次に定義の対象となる行番号を入力することになる。

5.2 エクセル形式のデータを読み込む場合

テキスト形式のデータの場合、次の内容をシンタックスエディタに記述することになる。

```
get data
    /type=xls
    /file='ファイルの場所¥ファイル名.xls'
    /sheet=name 'シート名'
    /cellrange=range'開始セル:終了セル'.
```

「/type」は、読み込むデータの形式を指定するサブコマンドである。エクセル形式のデータを読み込む場合は、「xls」と指定すればよい。「/file」は、読み込むファイルのある場所、読み込むファイル名を指定するサブコマンドである。「/sheet」は、データが入力されているシート名を指定するサブコマンドである。エクセルファイルを指定しても、ブック内のどのシートを読み込むか、ということまでを指定しなければならない。「/cellrange」は、データ（1 行目の変数名も含む）が入力されているセルの範囲を指定するサブコマンドである。「開始セル」にはエクセル上の左上端のセル番地（たいていは A1）を、「終了セル」は最終ケースの最後の変数が入っているセル番地を入力すればよい。例えば、終了セルが「IJ300」である場合、「/cellrange=range' A1: IJ300'.」となる。ちなみに、1 行目に入っている変数名は SPSS での変数名として登録される。

6　おわりに

シンタックスエディタに以上のコマンドを入力して実行すれば、SPSS のデータエディタにデータが読み込まれる。これが SPSS データファイルである。このファイルは忘れず保存しておく必要がある。以後、さまざまなシンタックスで分析を進めることになるが、その作業については他の文献（小林・雨森・山本 2007）を参照されたい。

〔文献〕

小林久高, 1993「SPSS 練習プログラム」小林久高編『同志社大学社会調査実習報告書』1:352-358, 同志社大学文学部社会学専攻.
――――, 1994「社会調査における入力データチェックプログラム」『社会学論集』1:165-178,奈良女子大学文学部社会学科.
――――, 1998「SPSS for Windows の使い方」『島根大学情報処理センター広報』9:39-50.
小林久高・雨森聡・山本圭三, 2007「SPSS を用いた社会調査データの分析―シンタックスの解説を中心に」小林久高編『同志社大学社会調査実習報告書　15』（第 1 分冊）:352-358, 同志社大学文学部社会学専攻.