９．データ・クリーニングのテクニック

ここでは、入力された素データのデータ・クリーニングに使う手法を紹介します。自前で調査をした場合に、最初にやらなければならない作業ですので、やや前倒し的に必要な操作だけ扱います。

手元にあるデータがすでにクリーニング済みの場合には、ここは読み飛ばして、１０章に進んでください。

9.1　データ・クリーニングとは

データ・クリーニングとは､入力されたばかりのデータの誤りをチェックする作業です。

通常、素データファイルが作成されるまでに、(1)（回答者自身または調査員による）調査票への記入、(2)記入された調査票の点検（エディティング)、(3)調査票に記載されたデータのコード化（コーディング)、(4)コーディングされたデータの入力、などの過程を経ています。

これらの過程で、さまざまな誤りが修正されると同時に、さまざまな誤りが修正されずに残ったり、あらたに誤りが生じてしまったりします。

たとえば、調査票への記入ミスは、エディティングの過程で発見され修正されるはずですが、こうしたチェックを免れて残っているミスが必ずといっていいほどあります。またコーディングの過程での転記ミスなどのコードの誤り、データ入力過程でのパンチミスなど、新たな誤りが生じる可能性もあります。

これらの誤りを発見し、可能な限り修正するのが、データ・クリーニングです。基本的には、ありえないコードの入った調査票を探し出し、調査票原票に戻って点検し、正しいコードに修正するという作業になります。

この作業の第一段階は、すべての変数について、度数分布表を出力し、ありえないコードが打ち出されていないかどうか、また、欠損値や非該当などの数に矛盾がないかどうかを調べることです。

第二段階は、ロジカル・チェックといい、論理的に関連のあるクロス表（たとえば、婚姻状態と子ども数､雇用状態と職業関連項目､非該当が出る枝分かれ質問など）を出力し、ありえないセルに該当するケースがないかどうかを調べることです。

どちらの場合にも、問題のあるケースを特定し、原票にあたって確認し、素データファイルを修正するのが原則です。

9.2　ID 番号の生成

ケースを特定するには、ケースごとに与えられている ID 番号が必要です。通常、原票には ID 番号が付けられています。しかし、ID 番号が、地点番号と組み合わされている場合には、地点データをひとつの変数として残しておくために、固有の ID が変数として準備されていない場合がしばしば生じます。

たとえば、４桁の ID のうち上１桁が地点番号だったとしましょう。そうすると上１桁は地点コードを示す変数、下３桁はサンプル番号として読ませることになります。本マニュアルで使用している例では、上１桁は chiten、下３桁は sample という変数名で ID を読

み込んでいます。地点は５つあるので、sample は同じ番号が最大５つあることになります。これでは、ケースを特定できません。

そこで、２つの変数を組み合わせて本来の４桁の ID を新変数として生成しておく必要があります。compute コマンドを使って、

```
compute ID=chiten*1000+sample.
```

とすることで、上１桁が地点番号、下３桁がサンプル番号のサンプル ID が生成されます。なお、compute コマンドは、既存の変数から新しい変数を生成するのにコマンドで、一般的な書式は、

compute 新変数名＝計算式.

です。計算式で使うことのできる加減乗除の演算記号は、加法＋、減法－、乗法＊、除法／で、乗・除は加・減に優先するので、通常の計算式を書くのと同じです。また、演算の優先関係を示す()は何重に用いてもかまいません。

compute コマンドは、計算式のなかで使う変数が定義されていれば、どの場所においてもかまいません。ID 番号は最も最初に生成する新変数なので、素データに存在する変数の値ラベルの後におくとよいでしょう。

9.3 度数分布表の作成

度数分布表は、どういう回答が何名あったかを知る最も基礎的な統計です。

一般的な書式は、

frequencies variables =変数名（あるいは変数リスト）.

freq 以下を省略して　**freq 変数名**　で通る。

例１　freq chiten.
例２　freq　q1 to q61.
例３　freq all.

データ・クリーニングでは、まずすべての変数の度数分布表を出力し、ありえないコードを発見したり、非該当のケース数からコーディング上の問題点を発見したりします。疑わしい場合を発見したら、9.5 の手法を使って、疑わしいケースを特定します。

9.4 クロス表の作成

ロジカル・チェックによってありえないコードを検出するには、クロス表を作成する必

要があります。データ・クリーニングに必要なのは度数だけのクロス集計（%の計算は不要）です。

一般的な書式は、

```
crosstabs tables 変数名 by 変数名.
```
→ cross 以下を省略して　**cross 変数名 by 変数名.**　でも通ります。

9.5　ある条件に該当する ID だけを打ち出す

　おかしなデータが見つかったら、その ID を特定する必要があります。そのためには、if 文を使って条件を絞る変数を作り、その変数を使って filter をかけて、その条件を満たす ID だけの度数分布表を出します。

たとえば、q1 が 1 で、q42 が 1 のサンプルの ID 一覧を作成するには、

```
if ((q1 eq 1) and (q42 eq 1)) check=1.
filter by check.
freq ID.
```

ここで、if 文の条件式に使うことのできる記号は、
= なら eq
＜なら lt
＞なら gt
≦なら le
≧なら ge
≠なら ne
である。

【filter を使って連続的に作業する場合の注意点】
①一度生成した変数 check は、たとえシンタックス・エディタ上のプログラムから削除しても、SPSS データ・ファイルには残っているので、引きつづき別の条件をかける作業をするときには、新しい変数名（たとえば check1）を使う必要があります。(いったん、SPSS データ・ファイルを捨てて、データを最初から読み込む場合は別です)。
②いったんかけた filter は、べつの filter がかかるまでは生き続けます。たんに filter を解除するには、

```
filter off.
```

というコマンドを実行します。

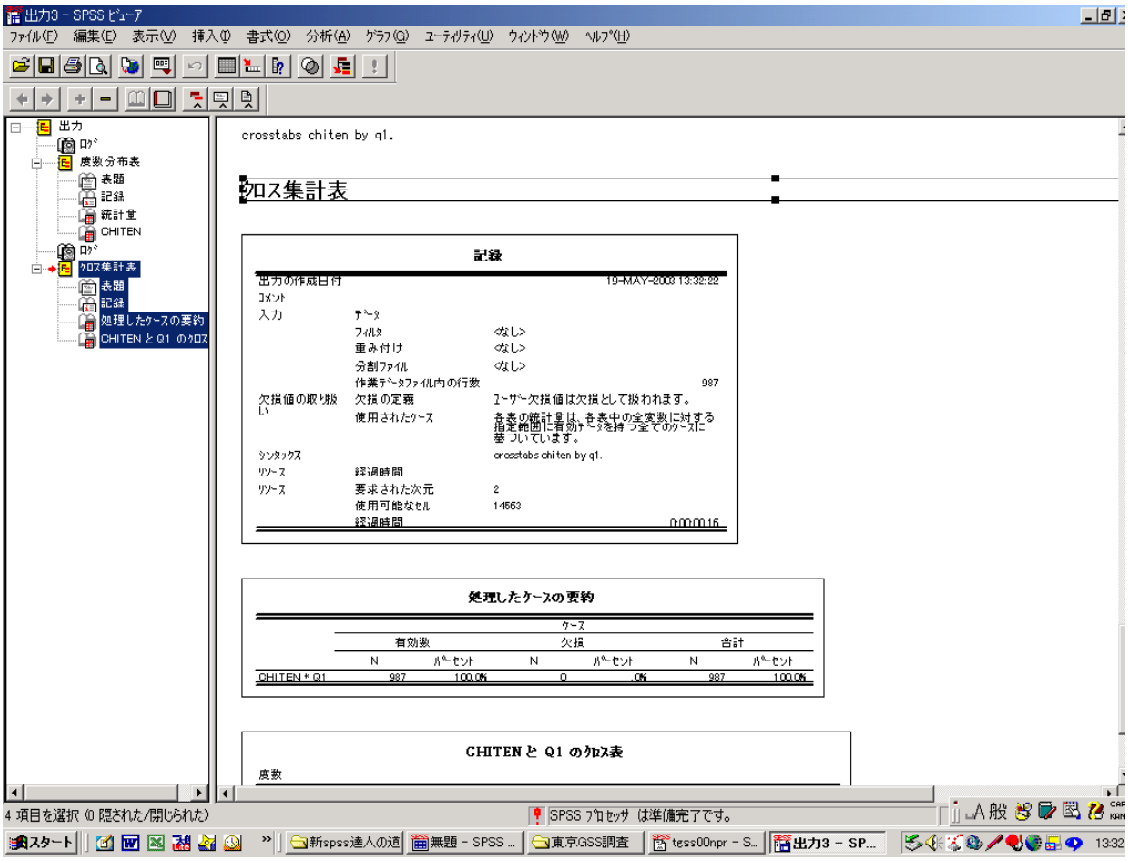
　出力ビューアに「記録」を出力する設定にしておけば、「記録」にどのような filter がか

かっているか（いないか）が表示されます。また、ログを出力する設定にしておけば、どのようなプログラムにもとづく結果であるかが表示されます。(→８．参照)。

9.6　出力結果の印刷

書力結果を印刷したいときには､印刷範囲を指定して(画面上でマウスをクリックする)、「ファイル」→「印刷」をプルダウンしていくか、あるいは印刷のアイコンをクリックすればできます。(図 9.1 参照)。

図 9.1 印刷範囲の指定



9.7　データの修正

このようにして検出されたエラーは、原票に戻って確認したうえで、素データ・ファイル上で修正します（機械的にプログラム上でコードを変更する場合には、if コマンドや compute コマンドを応用してください)。