

Conventions for naming variables in SPSS

Contributed by John F Hall
 Tuesday, 07 March 2006
 Last Updated Monday, 10 April 2006

Conventions for naming variables in SPSSTM(1)

John F Hall 30 March 2006 (2)

Whenever SPSS is called, it must have a file to be working on. This file is known as the active file and must have at least one variable in it. Imagine a small survey with questions about party support, age last birthday, sex and support for trade unions (on a five-point scale from 1: "make them illegal" to 5: "give my life to keep them in existence"). These are only a few of the questions asked and the data for each completed questionnaire fills more than one line. Thus we need to know:

- how the data for each variable has been coded
- which line it is on
- whereabouts in the line it is
- how many columns it occupies

Let us suppose that our variables have been coded as follows:

Variable number	Codes	Column(s)
Party support	Con, Lab, Lib-Dem, 1, 2, 3, 4, 5, 8, 9, Other, None, Refused	1, 2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100
Sex	M, F	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100
Age	16 - 90, 99	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100

Let us also suppose that the party support categories were hand-coded from their categories into numeric codes (1=Con, 2=Lab, 3=Lib-Dem, 4=Other, 8=Refused, 0=None), but that sex has been coded in its original alphabetic form. As well as the scale of support for trade unions, some people have refused to give an opinion, others have insisted on "Don't know": these have been coded respectively 8=Refused and 9=Don't know. Age has been recorded as the actual age last birthday in the range 16 to 90, but some people have not given their ages and these have been coded as 99.

Having decided which variables we want for our analysis we must first give them all names for our SPSS run. Variable names in SPSS can be any combination of up to 8 letters and numbers, but the first character must always be a letter of the alphabet. For beginners it may be easier to use a variable name which looks like plain English (e.g. SEX, AGE, PARTY, UNION) and these are known as mnemonic names (after the Greek for memory) and may be easier to remember.

1) SPSS is the registered trademark of SPSS Inc. Details of SPSS products and services are available from <http://www.spss.com> or from the UK site <http://www.spss.com/uk>

2) This document has been updated and edited by John Hall from an earlier version by himself and Jim Ring as applied to mainframe versions. It may be outdated by subsequent developments in SPSS, but the essential logic remains valid as do most of the procedures. The Windows version differs since it treats alpha data as string variables and, apart from initial reading of raw data, may not need to access or write external files as described here. Saved files from the data editor are now suffixed .SAV or .por, but syntax files still have .sps endings. Data capture has advanced by leaps and bounds with the arrival of portable computers which contain questionnaires, direct data entry and automatic quality controls. Manual data preparation is now obsolete for many large scale surveys, but is still used by individual students and researchers for their own projects. It is not clear whether the notion of a raw data matrix in 80-column format still has relevance now that data entry is made directly into packages such as SPSS or indirectly via spreadsheets such as Excel. This could in fact make errors much harder to spot in some instances.

{mospagebreak title=Page 2}

However, in a survey with a very large number of variables, several hundred names may be needed, and the generation of new and unique variable names, not to mention the ability to remember them all and which questions they relate to, constitutes a prodigious – and, to me, pointless effort which virtually guarantees errors and makes smooth working on large data files all but impossible.

The SPSS setup files compiled by Liverpool University and supplied via the UK Data Archive at Essex University for the SCPR British Social Attitudes series are all of this type and – there is a dedicated and forthright school of "Mnemonicism" among – the SPSS user community. SPSS files from the European Social Survey series also follow this naming convention.

– This can be confusing for teaching purposes (3) – and for research within a single year of the series, although possibly useful for cross-year comparisons as the same variable names are retained when questions are repeated in successive waves. – However, they cannot easily be used in conjunction with facsimiles of the original questionnaires (which have information for data preparation printed in the margins).

At both SSRC (4) – and PNL (5) – we developed a different approach to the naming of variables. – Because we have handled literally hundreds of questionnaire surveys and because we could not afford errors or lost time, and because we have had to deal with very large numbers of "naive" clients, many of whom had minimal documentation and resources other than their own time, we developed a convention which eschewed mnemonic names (except for a few standard variables such as SEX, AGE and MARITAL status) in favour of positional variable names.

The early versions of SPSS, as well as using mnemonic variable names, included a facility for automatic generation of variable names starting with the letters VAR followed by three digits (eg VAR001 – VAR002 etc). – By using the keyword TO between a pair of variable names (and provided the second name contained a number greater than that in the first – (eg VAR001 TO VAR051) SPSS automatically generated names for all the implied variables in between as well as the two named – (eg VAR001 VAR002 VAR050 – VAR051). – This enabled the specification of large numbers of variable names without having to write them all down one at a time. – Consequently many users gave the first variable in their data the name VAR001 and – produced automatically generated names for each variable until they got to the – last one. – Thus a survey with 240 variables would have had variables specified as VAR001 TO VAR240, which is a lot – quicker and easier than typing out 240 separate variable names.

We dub this convention sequential naming of variables.

3) At PNL we changed the names of the variables (using the SPSS RENAME – command) to comply with the following convention, which, in conjunction with copies of the original questionnaires, everyone finds much easier to follow and use.

4) Social Science Research Council Survey Unit: the authors were respectively Senior Research Fellow and Research Officer

5) Polytechnic of North London – Survey Research Unit: the authors were respectively Unit Director and Senior Research Officer

{mospagebreak title=Page 3}

Later versions of SPSS allowed more flexible naming of this kind: any (combination of) letter(s) could be used at the beginning of the name and the requirement for 3 digits was dropped. – Some researchers immediately started calling their variables V1 TO V69 or Q1 TO Q69 and the like, which was certainly an improvement as the names at least bore a more transparent relationship to the questionnaire from which they were purportedly derived! – Also the ability to use Vddd instead of VARddd saved a lot of key depressions when specifying analyses.

However even this can get cumbersome once we get questionnaires with large classification sections (i.e. no question numbers) and questions with multiple subquestions. – So far SPSS – does not offer automatic generation of names for variables whose sequence is defined by alphabetical order of their last letter (e.g. Q21a – Q21b – Q21c etc).

The SSRC Survey Unit was attached to the London School of Economics and through the LSE we had access to the CDC2000 at the University of London Computer Centre. – It was here (in the early 70's) that we had our first taste of SPSS. – Before SPSS we relied on commercial services (eg Donovan Data Systems at Research Services Ltd) which could produce client friendly tables (with labels and full – text) or we used LSE's own survey data tabulation program, SDTAB (6).

The latter was limited in its facilities, but had a convention of naming variables by their position on an 80-column Hollerith card. – SDTAB variables could only consist of a three digit number in which the first digit indicated which card, and the second and third digits which column, the data were to be read from. – Thus 151 would be the name of the variable on column 51 of card 1 and 237 would be on card 2, column 37. – Remember in – those – days everything, – including SPSS

programs, had to be punched on cards and read through a card reader. (Prior to that even the tabulations were performed on a card sorter.) With this program, data was limited to 9 cards per survey.

SDTAB's variable naming had the distinct advantage that computer printout could instantly be related to the data layout for a questionnaire, and vice-versa, provided, of course, that the questionnaire had been properly laid out, not just to record respondents' answers, but also for use as a data-entry tool.

What we started doing at the SSRC Survey Unit, and developed further at PNL, was to combine the automatic name generating capacity of SPSS with the variable naming logic of SDTAB. We also made one major departure from sequential naming. In our convention the numeric part of the variable name relates, not to its sequential position in the questionnaire, but to its actual location in the coded data. This means that names immediately relate to the line of data within each case, and to the position, or column(s), within each line. In all our variable names, the last two digits tell us which column(s) in a data line contain(s) the coded data for the variable, and the digit (or digits) before them tells us which line of data the variable is on.

Thus, to us, V106 indicates the variable which is coded on line 1, column 6 of the data (It does not mean the 106th variable in the file!). Conversely, if we have to name a variable, the data for which is to be entered on line 3 in column 44, we should call it V344. Ah, you may say. But what if it's a variable like age which you have coded as two digits on line 3, columns 14 -15? Well, when a variable is spread across more than one column (in computer jargon, a "field" more than one column wide) we always call the variable by the first column in the field. In our example, age would be called V314 and there would not be a variable V315. The next variable after age would then be V316.

All variables in the final version of the SPSS saved system file will have names beginning with a letter V followed by three digits (or four (7) if there are ten or more records per case). Exceptions will be derived variables such as scores from attitude scales or complex derived variables such as income when it has been computed from several other variables. Even the standard classification variables such as sex, age and income groups will start life in V-form.

This positional convention is much preferable to sequential naming for one other important reason. Most decent questionnaires already have printed on them information for the data-preparation people to punch data on to 80 column cards or, more likely these days, keying the data directly into the computer. Either way, the data will appear in the computer as lines in such a way that the coded data for each question appear on the line(s) and in the position(s) indicated by the layout information for data-prep which normally appears down the margins of the questionnaire pages.

This gives the immediate and enormous advantage that the original questionnaire can be used as a document for understanding what variables are called and what they are. Thus V528 can immediately be checked by looking for the question that has been coded on line 5 column 28: conversely the question that has been coded on column 28 on line 5 can be expected to be called V528 (at least on most surveys handled by the Survey Research Unit at PNL)

To go back to our original example:

Variable	Codes	Line number
Party support	Con, Lab, Lib-Dem, 1, 2, 3, 4, 5, 8, 9, Other, None, Refused	1
Union support	1, 2, 3, 4, 5, 8, 9	2
Sex	M, F	
Age	16 - 90, 99	

These data have, let's say, been stored on drive A: in a data file called POLY.DAT and there are three data lines per case. We want to read in the data for these four variables and then save them on drive A: in a file called POLY.SAV for future use. Remember, we have coded party support using numeric codes, but left sex with its original alphabetic codes.

How do we get SPSS to read these data and save them in a saved system file?

6) SDTAB was written by Peter Wakeford, then Director of Computer Services at LSE

7) For example, the data for the 1989 British Social Attitudes survey, used in the PNL course Survey Analysis Workshop, has 23 records per case.

{mospagebreak title=Page 4}

Let's take it one step at a time.

Step 1:

Give each of your variables a name. (Just this once you can use mnemonics!) The name must start with a letter of the alphabet, but must not be longer than eight characters in total. Thus: PARTY UNION SEX AGE (In PNL convention these would have been V127 V234 V307 V308)

Step 2

Work out for each variable whether it has been coded in alphabetic format or numeric. If it is numeric, does it have any decimal places? Very few surveys ever collect data with decimals unless they are financial or medical, so we don't need to bother ourselves with that just yet. We do need to specify for each variable what TYPE it is, "alpha" (A) or "integer" numeric (0) with no decimal places.

Step 3

For each variable, work out which record (line of data) it is on and which column(s) it occupies in the line.

Step 4

To specify variables in SPSS we have to start with the line on which the first variable occurs. We specify the line we want by writing a forward slash "/" followed by the number of the line: this is line 1, so we start by:

/1

Step 5

Next we write down the name of the first variable we want from that line. There is only one and we want to call it PARTY.

/1 PARTY

Step 6

Write in the column(s) within the line containing the coded data for the variable.

/1 PARTY 27

Step 7

Finally, write in brackets the type of data coded for the variable (in this case, an integer number with no decimals)

/1 PARTY 27 (0)

Well, that's our first variable specified: now let's do the rest.

A little table helps to show us what to do:

Line	Variable name	Position	Type
1	PARTY	27	(0)
2	UNION	34	(0)
3	SEX	7	(A)
3	AGE	8-9	(0)

So, going over all the steps again we get: /1 PARTY 27 (0)
 /2 UNION 34 (0)
 /3 SEX 7 (A)
 /3 AGE 8-9 (0)

Note that if you have more than one variable on the same data line you only need to give the line number once.

What we have just done is to specify a DATA LIST exactly as SPSS needs it before it can read raw data. Before we can use it, however, we need to give SPSS some more information. We need to tell SPSS which file (8) the data are to be read from and also how many records (lines of data) there are for each case.

We do this by means of the DATA LIST command, followed by two subcommands, FILE = and RECORDS = ("records" is a SPSS-ese for "lines" or "cards")

eg DATA LIST FILE = < name of external data file >
RECORDS = < number of records per case >

"record" in SPSS can be a line of up to 255 columns, but in practice we tend to use 80, partly because that is how many columns there are on a standard Hollerith card, but mainly because that is the number of columns on a standard "line" on a computer terminal with a visual display unit (VDU).

In this case we have 3 records per case, so RECORDS = 3 will be needed. But what do we do about FILE = ? You will remember that every SPSS job needs to have at least one file it can access from somewhere else in the computer. This is known as an external file. To use such a file, SPSS needs to be given the (location and) name of the external file enclosed in single primes (eg 'A:POLY.DAT').

So here we go! DATA LIST FILE = 'A:POLY.DAT'

```
RECORDS=3
/1 PARTY 27
/2 UNION 3
/3 SEX AGE
```

SPSS will not actually do anything at this stage (unless you override it, but that's later). In order to start reading the data, we must first ask SPSS to do something such as performing a statistical analysis or listing the variables or part of the data.

We can now ask for frequency tables or some other statistical analysis. For frequencies we could write:

```
FREQUENCIES VARIABLES = PARTY UNION SEX AGE
```

Alternatively, we can simply SAVE the work we have done and store it our own private work area. To save the active file and store it safely away for future use we write:

```
SAVE (9) OUTFILE = 'A:POLY.SAV'
```

Oh dear! Now what? Well, it's back to the old "external file" again, but this time it's going to be a saved file with a .SAV extension on drive A: which we shall call POLY.SAV (the file name needs to be enclosed in single primes).

```
SAVE OUTFILE = 'A:POLY.SAV'
```

The full program then looks like this.... DATA LIST FILE = 'A:POLY.DAT'

```
RECORDS = 3
/1 PARTY
/2 UNION
/3 SEX AGE
FREQUENCIES VARIABLES = PARTY UNION SEX AGE
SAVE OUTFILE = 'A:POLY.SAV'
```

Setting up DATA LIST for very large files is both time consuming and prone to error, especially for beginners. It's also a job which is sometimes best left to experts, although there are some researchers who insist on trying to do everything themselves, even if it takes them ten times as long (and costs ten times as much!). However, even for them, we suggest that our positional variable naming will save time and money (and tears!) as will be evident when you check the output from an SPSS DATA LIST command on which any errors of record or column location will be immediately apparent as the variable names are printed out side by side with the record and column(s) from which they have been read (See example on next page).

There's a lot more to setting up SPSS saved system files, such as specifying MISSING VALUES, VARIABLE LABELS, VALUE LABELS, using RECODE to change any or all of your alpha variables into numeric or to group variables into

