

Block 2: Analysing one variable

2.3 Data transformations

2.3.1.1 Data transformations

[29 November 2010: updated 21 June 2013]

This is where SPSS, with its powerful procedures for data management and intuitive English-like language, really comes into its own.

These features (unique at the time) account for the meteoric spread of SPSS in the 1970s and becoming the industry standard in universities, colleges and (central and local) government for the processing, management and analysis of data from questionnaire surveys and administrative sources.

In SPSS we can:

1: Select **cases** for analysis page 2

[**SELECT IF** ~ ~ ~ ~]

2: Select **variables** for analysis page 5

[**GET /KEEP** ~ ~ ~ ~]

3: Change the **names** of variables page 7

[**RENAME VARIABLES** ~ ~ ~ ~]

4: Change or group the **values** of variables page 11

[**RECODE** ~ ~ ~ ~]

5: Create **new variables** by

a: Recoding values page 11

[**RECODE** ~ ~ ~ ~ **INTO** ~ ~ ~ ~]
(eg from **age** into **agegroup**) :

b: Performing calculations based on the values of one or more variables.

[**COUNT** ~ ~ ~ ~]

[**COMPUTE** ~ ~ ~ ~]

[see 3.5.1

[An introduction to COUNT and COMPUTE](#)]

[**IF** ~ ~ ~ ~]

[**DO IF** ~ ~ ~ ~ **ELSE IF** ~ ~ ~ ~]

[see 3.4.1

[Tutorial - Conditional transformations](#)]

These operations are called **data transformations**.

1: Selecting cases for analysis

We can select **cases** on the basis of their values on one or more variables by using the **SELECT IF** command followed by a **logical expression**.

Format:

```
SELECT IF ( <logical expression1> )
```

To select the women from a sample (say, value **2** on variable **sex**) and perform subsequent analyses on women only, we write:

```
select if (sex eq 2) .
```

This selection remains in force throughout the current session: all subsequent procedures will be for women only. If you save your working file while the selection is in force, you will permanently lose all cases except the women. This is why you should always have **at least one read only copy** (and preferably two) of your original data and saved files stored safely away from your computer on a CD or other medium such as tape or a remote server.

If we only want a **temporary** selection for a single procedure we simply put a **TEMPORARY** command before the **SELECT IF** command:

```
eg temporary .  
select if (sex eq 2) .
```

This command selects women for the next statistical procedure only, after which the data set reverts to the full sample for the next procedure.

SELECT IF can be useful in helping to understand the logic of survey analysis. The word "analysis" is derived from an Ancient Greek verb meaning "to break up". We can calculate a statistic for a whole sample (eg mean life-satisfaction on a 0-10 scale; percentage "very happy") and then use **SELECT IF** to break it down into its constituent parts to demonstrate that the overall figure for the mean or percentage in the sample is a "weighted average" of the means and percentages within the subsamples, which in turn are "weighted averages" of the means and percentages in sub-sub-samples and so on.

This notion is useful as a transition from working with one variable to working with two or more variables since, although **SELECT IF** implies the use of another variable, we are still only analysing one variable at a time. For students new to survey analysis and statistical ideas, this makes the progression from frequency counts to two-way and three-way contingency tables easier to understand and follow.

Several temporary selections may follow each other in a single session. For instance, if we want to look at the relative rankings of the topic "Welfare State" (**v4**) in Q.1 of the pre-course questionnaire, first for the whole sample, then separately for men only and women only:

¹ SPSS evaluates a logical expression as either TRUE or FALSE and performs subsequent analyses only if the expression is evaluated as TRUE. It can contain logical symbols for:

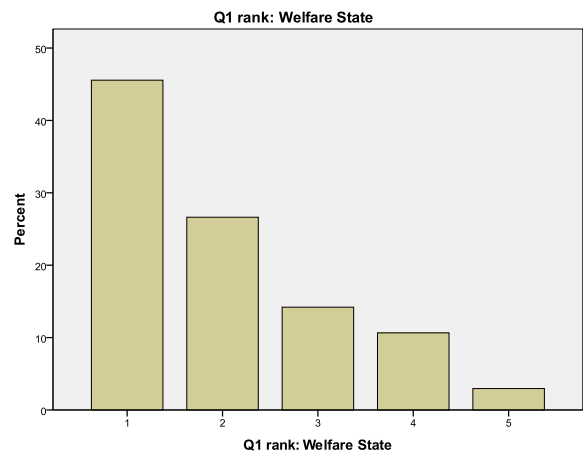
| | | | | | |
|--------------------------|----|----|-----------------------|----|----|
| equal to | = | EQ | not equal to | <> | NE |
| greater than | > | GT | less than | < | LT |
| greater than or equal to | >= | GE | less than or equal to | <= | LE |

. . . sometimes with combinations using logical operators **AND**, **NOT** or **OR**, or SPSS reserved keywords **MISSING**, **SYSMIS** or **VALUE**. Use of brackets is advisable and usually essential. It's a bit early in the course to deal with here, but detailed explanations and examples will be given as and when appropriate. There's a very good explanation on pp 154 ff in Sarah Boslaugh's book (See [SPSS Textbooks](#) on this site)

Subtitle 'Whole sample' .
 frequencies v4 /barchart .

v4 Q1 rank: Welfare State

| | Frequency | Percent | Valid Percent | Cumulative Percent |
|---------|-----------|---------|---------------|--------------------|
| Valid 1 | 77 | 45.6 | 45.6 | 45.6 |
| 2 | 45 | 26.6 | 26.6 | 72.2 |
| 3 | 24 | 14.2 | 14.2 | 86.4 |
| 4 | 18 | 10.7 | 10.7 | 97.0 |
| 5 | 5 | 3.0 | 3.0 | 100.0 |
| Total | 169 | 100.0 | 100.0 | |



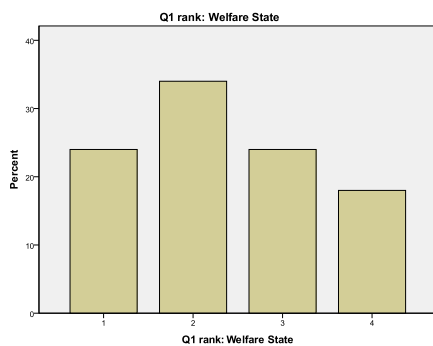
These results can be "broken down" into **conditional** frequency tables and charts demonstrating that the overall figure of 45.6% ranking "Welfare State" as the most interesting topic (the **dependent** or **criterion** variable) can be broken down into two **conditional** figures of 24% for men and 54.6% for women when **controlling** for sex (the **independent** variable). This concept of a **control** or **test** variable underlies much of the logic of survey analysis and, eventually, statistical modelling.

Subtitle 'Men only' .
 temporary .
 select if (sex eq 1) .
 frequencies v4 /barchart .

Men only

v4 Q1 rank: Welfare State

| | Freq | Percent | Valid Percent | Cumulative Percent |
|---------|------|---------|---------------|--------------------|
| Valid 1 | 12 | 24.0 | 24.0 | 24.0 |
| 2 | 17 | 34.0 | 34.0 | 58.0 |
| 3 | 12 | 24.0 | 24.0 | 82.0 |
| 4 | 9 | 18.0 | 18.0 | 100.0 |
| 5 | 0 | 0 | 0 | 100.0 |
| Total | 50 | 100.0 | 100.0 | |



Men only

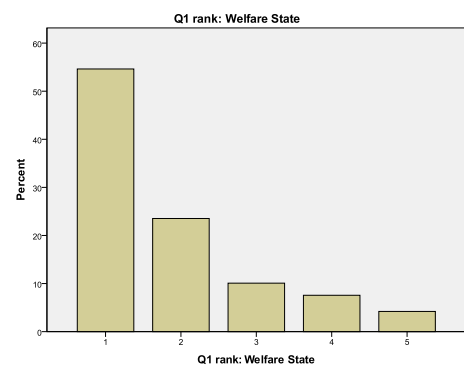
[NB: There are no men with value 5: I've inserted the missing figures in red]

Subtitle 'Women only' .
 temporary .
 select if (sex eq 2) .
 frequencies v4 /barchart .

Women only

v4 Q1 rank: Welfare State

| | Freq | Percent | Valid Percent | Cumulative Percent |
|---------|------|---------|---------------|--------------------|
| Valid 1 | 65 | 54.6 | 54.6 | 54.6 |
| 2 | 28 | 23.5 | 23.5 | 78.2 |
| 3 | 12 | 10.1 | 10.1 | 88.2 |
| 4 | 9 | 7.6 | 7.6 | 95.8 |
| 5 | 5 | 4.2 | 4.2 | 100.0 |
| Total | 119 | 100.0 | 100.0 | |



Women only

If we want to look at women of pensionable age in the general population (currently 60 in UK) this involves selecting not just for sex, but also for age , and we might write:

eg `select if ((sex eq 2) and (age ge 60)) .`

SPSS also recognises mathematical symbols in these expressions:

eg `select if ((sex = 2) and (age >= 60)) .`

Note the use of double brackets in the above. Although SPSS will sometimes work without the brackets, it is best to stick to them for clarity and logical precision.

Variables with alpha values require the use of primes to specify those values:

eg `select if (sex = 'F') .`

Some of these logical expressions can get highly complex, but that's plenty to be going on with.

2: Selecting variables for analysis

Just as it is possible to select **cases** for analysis by the use of **SELECT IF**, so it is also possible to select **variables** for analysis. Modern personal computers are very fast and have enormous storage and memory functions, but researchers using networks or campus mainframes may be rationed as to disk space. Any attempt to download a large SPSS saved file is like trying to empty a milk tanker into an egg-cup and could cause an immediate overflow.

Even with the speed and storage capacity of modern computers and PC's it is useful to limit the size of the working file by reducing the number of variables used, especially if there is a very large number of cases. This will at least save some processing time as well as making the contents of your file easier to see in the screenshots.

Accordingly it is useful, and sometimes necessary, to limit the size of the working file. This can be done using **GET FILE** with **/KEEP**.

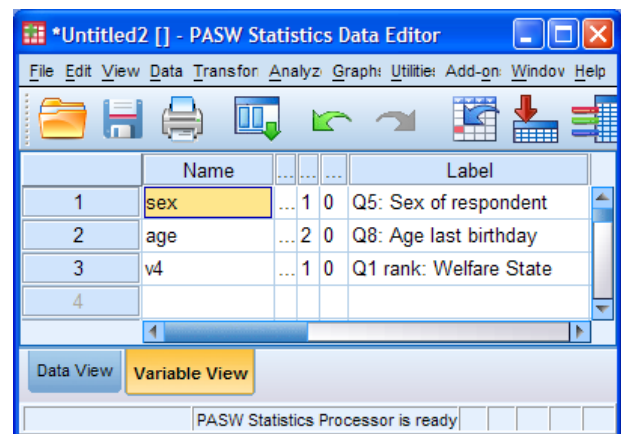
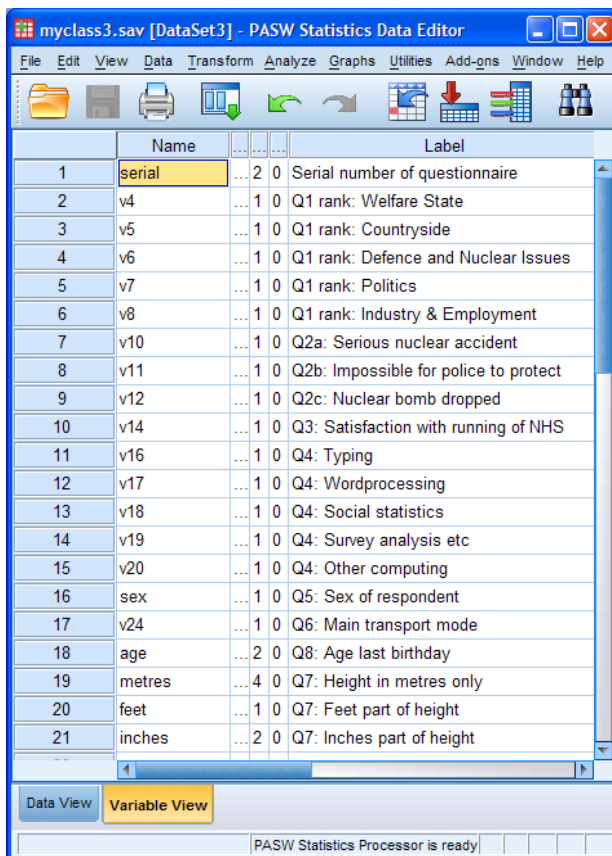
Format:

```
GET
FILE '<filename>.sav' >
/KEEP <varlist> .
```

[NB: Variables are selected in the order specified in <varlist> not in their original file order]

eg **GET**
FILE 'C:\Documents and Settings\Owner\Desktop\myclass\myclass3.sav'
/keep sex age v4 .

... creates a data editor containing only three variables, and in a different order:



Original using **GET FILE** ~ ~ ~ only . .

... and after using **GET FILE** ~ ~ ~
/keep sex age v4 .

3: Changing the names of variables

RENAME VARIABLES changes the name(s) of existing variables.

General format

RENAME VARIABLES (<oldvarlist> = <newvarlist>).

eg **rename variables** (v1411 v1412 = sex age).
or **rename variables** (v1411 = sex) (v1412 = age) .

The renamed variables **sex** and **age** replace **v1411** and **v1412** and will be in the same positions in the file, but will retain missing values and variable and value labels. This will be a permanent change if you save the working file.

[NB: If you want to keep both the original and the new variables use **COMPUTE** .

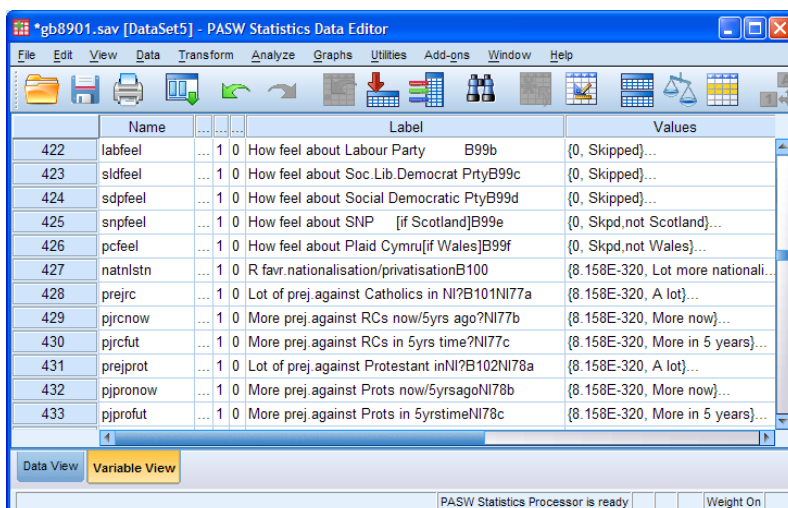
eg **compute sex = v1411** .
compute age = v1412 .

The computed variables **sex** and **age** will be added to the end of the file, but with no missing values or variable and value labels.]

We have already seen how to read raw data into SPSS from an external file using both **mnemonic** (eg **SEX** or **AGE**) and **positional** variable names² (eg **V1411 V1412** etc.). In files with a very large number of variables it is preferable to use names which relate directly to the data layout for the original questionnaire or, if there is no data layout, directly to the question numbers (eg **Q2 Q4a Q4b** etc.) With the questionnaire to hand (often the only documentation) this will make the data easier to define and the variables much easier to find. In a file with over 1,000 variables how do you find a variable called **PTYSPT**? Admittedly you can re-arrange all the variables in alphabetical order, but that makes it virtually impossible to work from the questionnaire or to use the **TO** convention.

The original SPSS saved files (created by John Curtice et al., Strathclyde University) for the 1986 and 1989 [British Social Attitudes](#) surveys have variables with names like **wwrelchd** and **marview**. These names may be useful for analysis by the investigating teams who designed the original questions and are familiar with the data, or who prefer the names to be retained across survey waves to make trend analysis easier, but where are they in the file?

Here's a screen-shot of part of the original 1989 file as supplied by the UK Data Archive, using **mnemonic** variable names, cluttered variable labels and some strange value labels.



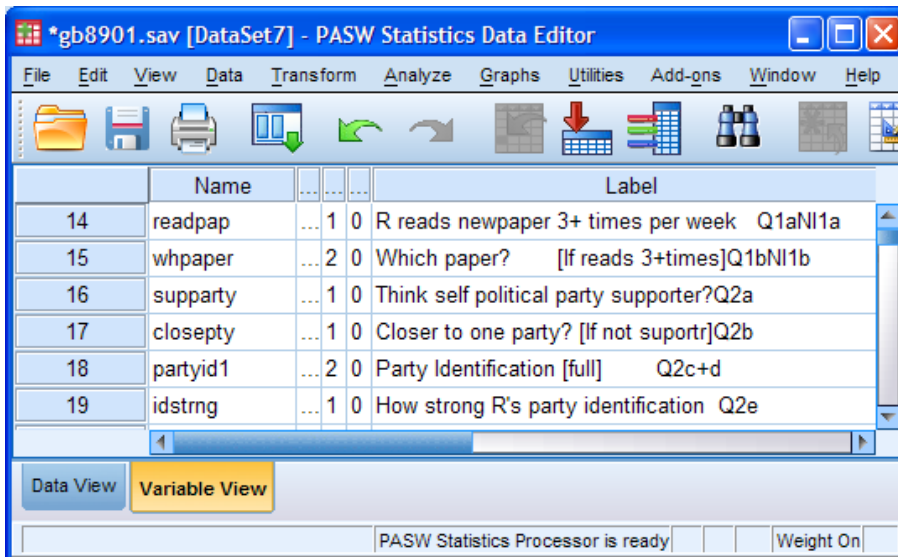
| | Name | ... | ... | Label | Values |
|-----|----------|-----|-----|---|-------------------------------------|
| 422 | labfeel | ... | 1 0 | How feel about Labour Party B99b | {0, Skipped}... |
| 423 | slffeel | ... | 1 0 | How feel about Soc.Lib.Democrat PrtyB99c | {0, Skipped}... |
| 424 | sdpfeel | ... | 1 0 | How feel about Social Democratic PtyB99d | {0, Skipped}... |
| 425 | snpfeel | ... | 1 0 | How feel about SNP [if Scotland]B99e | {0, Skpd_not Scotland}... |
| 426 | pcfeel | ... | 1 0 | How feel about Plaid Cymru[if Wales]B99f | {0, Skpd_not Wales}... |
| 427 | natnlstn | ... | 1 0 | R favr.nationalisation/privatisationB100 | {8.158E-320, Lot more nationali...} |
| 428 | prejrc | ... | 1 0 | Lot of prej.against Catholics in NI?B101NI77a | {8.158E-320, A lot}... |
| 429 | pjrcnow | ... | 1 0 | More prej.against RCs now/5yrs ago?NI77b | {8.158E-320, More now}... |
| 430 | pjrcfut | ... | 1 0 | More prej.against RCs in 5yrs time?NI77c | {8.158E-320, More in 5 years}... |
| 431 | prejprot | ... | 1 0 | Lot of prej.against Protestant inNI?B102NI78a | {8.158E-320, A lot}... |
| 432 | pjpronow | ... | 1 0 | More prej.against Prots now/5yrsagoNI78b | {8.158E-320, More now}... |
| 433 | pjprofut | ... | 1 0 | More prej.against Prots in 5yrstimeNI78c | {8.158E-320, More in 5 years}... |

² For an explanation of positional and mnemonic names, see 1.3.1 [Conventions for Naming Variables in SPSS](#)

Here's a facsimile of the first page of the 1989 British Social Attitudes questionnaire:

| | | Col./ Code | Skip to |
|--------------------|--|---|--|
| - 1 - | | | |
| SECTION ONE | | | |
| 1.a) | Do you normally read any daily <u>morning</u> newspaper at least 3 times a week? | Yes No | 216 1 + 2 + b) Q.2 |
| | IF YES | | 217-18 |
| b) | Which one do you normally read? IF MORE THAN ONE ASK: Which one do you read <u>most</u> frequently? | (Scottish) Daily Express Daily Mail Daily Mirror/Record Daily Star The Sun Today Daily Telegraph Financial Times The Guardian The Independent The Times Morning Star | 01 02 03 04 05 06 07 08 09 10 11 12 |
| | Other Irish/Northern Irish/Scottish/regional or local <u>daily morning</u> paper (WRITE IN) _____ | | 94 |
| | Other (WRITE IN) _____ | | 95 |
| 2.a) | ASK ALL Generally speaking, do you think of yourself as a supporter of any one political party? | Yes No | 219 1 + 2 + d) b) |
| | IF NO AT a) | | 220 |
| b) | Do you think of yourself as a little closer to one political party than to the others? | Yes No | 1 + 2 + d) c) |
| | IF NO AT b) | | |
| c) | If there were a general election tomorrow, which political party do you think you would be most likely to support? CODE ONE ONLY UNDER COL c) & d) IF ALLIANCE, PROBE: Social and Liberal Democrat or SDP (Owen)? | Conservative | c & d 221-22 |
| | IF YES AT a) OR b) | Labour | 01 |
| d) | Which one? CODE ONE ONLY UNDER c) & d) IF ALLIANCE, PROBE: Social and Liberal Democrat or SDP (Owen)? | Social and Liberal Democrat/Liberal/SLD SDP/Social Democrat Alliance (AFTER PROBE) Scottish Nationalist Plaid Cymru | 02 03 04 05 06 07 e) |
| | Other party (WRITE IN) _____ | | 08 |
| | Other answer (WRITE IN) _____ | | 09 |
| | None | | 10 |
| | Refused/unwilling to say | | 97 Q.3 |
| | IF ANY PARTY CODED AT c) & d) | | 223 |
| e) | Would you call yourself very strong ... (QUOTE PARTY NAMED) ... fairly strong, or not very strong? | Very strong Fairly strong Not very strong Don't know | 1 2 3 8 |

. . . and here is the relevant section of the original SPSS saved file relating to the questions. This one's a bit easier to interpret:



. . . but the variable labels could be improved by dropping the Northern Ireland (NI) part and moving the question number to the beginning of the label.

For use by others on a single wave of the survey, eg for teaching, it's much easier to use positional names like **v216 to v223** which relate directly to the data layout on the above questionnaire and to the original raw data layout in 80-column format. For our purposes we need a modified file with **mnemonic** variable names changed to **positional** using the **RENAME VARIABLES** command.

For the variables in the above screenshot:

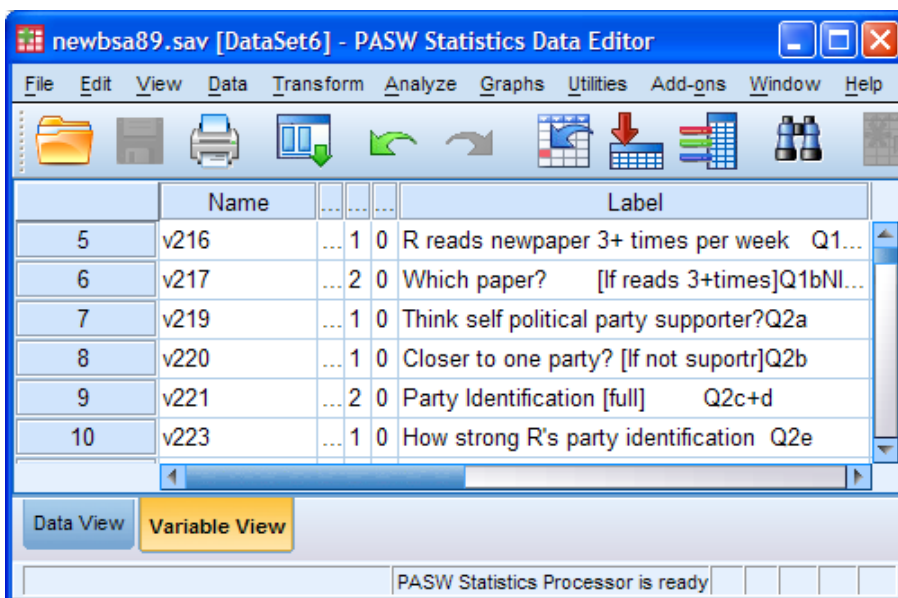
rename variables

(readpap=v216) (whpaper=v217) (supparty=v219)

(closepty=v220) (partyid1=v221)

~ ~ ~ ~

[continues for all variables in the file except agency-supplied derived variables]

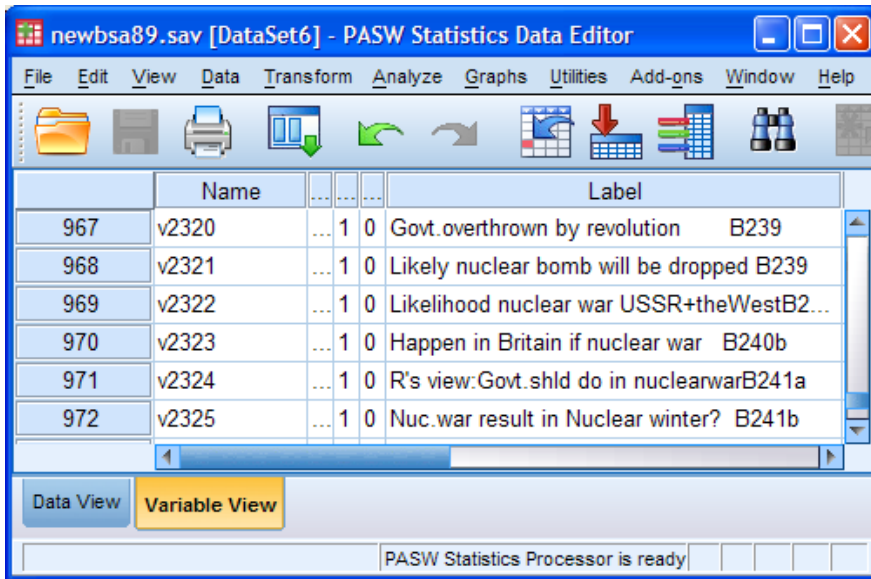


~ ~ ~ ~

(protrcmx to ctryjobs=v2245 to v2279)

(country1 to farmers=v2308 to v2312)

(predict1 to nucwintr=v2315 to v2325).



In surveys which no longer use 80-column formats, but use direct data capture systems such as CAPI, CATI or BLAISE (eg more recent waves of British Social Attitudes or the [European Social Survey](#)) there may well be no data layout information on the questionnaire at all. In such cases, it is easier to use question numbers.

[See [Old Dog, Old Tricks: Using SPSS Syntax to Avoid the Mouse trap](#) (pp 19 - 30) also the accompanying slide-show [SPSS usage in major surveys](#) for my critical examination of SPSS usage in the British Social Attitudes survey and the European Social Survey, together with my suggested improvements.]

4: Changing the values of variables

Variables with many values (eg **age last birthday** or **annual income** in ££ or \$\$) need to be grouped for tabulation with **CROSSTABS**. Other variables may have some categories with very few cases and the categories will need to be grouped with other appropriate values (eg **V11** in **myclass3.sav** has only 9 cases with value 4 and should be grouped with value 3 for cross-tabulation). Some items in rating scales may also need to be (at least temporarily) reversed before scale scores can be calculated.

RECODE changes one or more **values** of a variable (or set of variables) into a **new value**. The new value(s) can be written to the **existing** variable or written to a **new** variable (the old variable is retained with its original values).

General format:

```
RECODE <varlist> ( <value list1> = <target value1> )
                ( <value list2> = <target value2> )
                ~ ~ ~ ~
                [ INTO <target varlist> ]
                / <varlist> ~ ~ ~
```

where <value list> may consist of one or more of the following, separated by blanks or commas:

```
number
number THRU number
LOWEST THRU number
number THRU HIGHEST
MISSING (for user-missing values)
SYSMIS (for system missing value)
ELSE (any other value not previously included)
(CONVERT) (for conversion from alphabetic printing digits to numeric)
```

Examples:

```
recode v11 (4 = 3) .
recode v2018 to v2023 (1=2) (5=4) .
recode age (18 thru 29 = 1) (30 thru 44 = 2)
           (45 thru 59 = 3) (60 thru 97 = 4)
           (else=sysmis) into agegroup .
```

Alphabetic values must be in single primes. Variables with alphabetic values recoded to numeric must be written **INTO** a new variable:

```
recode a3 ('M' = 1) ('F' = 2) into sex .
recode ascale ('0'=0) ('1'=1) ('2'=2) ('3'=3) ('4'=4)
              ('5'=5) ('6'=6) ('7'=7) ('8'=8) ('9'=9)
              ('-'=11) ('+'=12) into scale .
recode ascale ('-'=11)('+'=12) (convert)
              into scale .
```

The last two examples both convert the printing characters '0' to '9' to their internal numeric values of 0 to 9, but letters and special characters need to be specified individually.

As SPSS encounters each **RECODE** command, it checks each value list in turn. If the current case has a value within the range specified in the value list, that value is recoded to the specified target value. If no value list covers the value, it is left unchanged. Unless they are specifically recoded to another value, user-missing values will be recoded to **SYSMIS**. Only one target value may be defined for each value list.

If you create new variables by grouping the values of an existing variable, it's good practice to specify variable and value labels at the same time. If the new variable is integer you can specify a **format** inside syntax or change it later in the data editor.

```
eg  recode
      age   (18 thru 29 = 1)(30 thru 44 =2)
           (45 thru 59 = 3)(60 thru 97 =4)
           (else=sysmis) into agegroup .
```

```
      variable labels
      agegroup 'Age group of respondent' .
```

```
      value labels
      agegroup   1 '18 - 29'
                  2 '30 - 44'
                  3 '45 - 59'
                  4 '60 or over' .
```

Be very careful with recodes! If you save the working file, **your changes will be permanent!**

As with all data transformations, it's sometimes safer to use the **temporary** command.

```
eg  temporary .
      recode ~ ~ ~ ~ .
```

You should always keep at least one read-only copy of your original file on a removable medium or on a safe server.

As with most procedures in SPSS, all the above data transformations can be done with the drop-down menus, but it's much quicker, easier and clearer with syntax. You're welcome to try the menus yourself, but you are advised to **PASTE** the syntax from the menus (or from the output file) into a separate syntax editor to save and keep for use with **copy** and **paste** in future analyses.

Now go to [2.3.1.2a1 Select and rename variables](#)

[\[Back to Block 2 menu\]](#)