

Session 2

Crosstabulation and Recode

	<i>page</i>
Missing Data	2-2
Crosstabulation in STATA	2-5
Recoding	2-7
Another way to recode	2-12
Computing New Variables	2-14
Example	2-15
Selecting Cases	2-16
Sampling Cases	2-19
Split Analysis	2-20
Practical Session 2	2-24

SESSION 2: Missing Data

STATA has 27 numeric missing values. The system missing value, which is the default missing value is '.'. However, **STATA** has the 'extended missing values'. These are defined as .a, .b, .c, ..., .z. Numeric missing values are represented by large positive values. The ordering is

all non missing numbers < . < .a < .b < ... < .z

When we have missing data, we have to be careful on the selection statement. If we use the expression `age > 30`, then all ages greater than 30 will be selected, as well as all missing ages.

To exclude missing values ask whether the value is less than ".". For instance,

```
. list if age > 30 & age < .
```

STATA has one string missing value, which is denoted by "".

When inputting data, codes representing information not collected or not applicable (e.g. the code 99 for age, meaning 'No response') need to be specified as missing. This is done by giving these codes a '*letter*'. This will cause **STATA** to omit respondents with these values from calculations (it would not be correct to calculate the average age of the sample including 99 as a valid value since that value does not mean that the respondent is 99 years old, but that no information on age was collected for that individual).

- Open the file '*example.dta*'.
- Open the **STATA** Data Editor.
- Create a new observation, leaving the value for **age** blank.

If we list the data, we would have something like the following window.

```
. list, separator(6)
```

	id	age	sex	var1	var2	var3
1.	1	22	M	4	2	1
2.	2	40	F	2	3	1
3.	3	27	M	3	3	2
4.	4	35	M	2	2	4
5.	5	24	F	1	2	2
6.	6	.	F	1	2	3

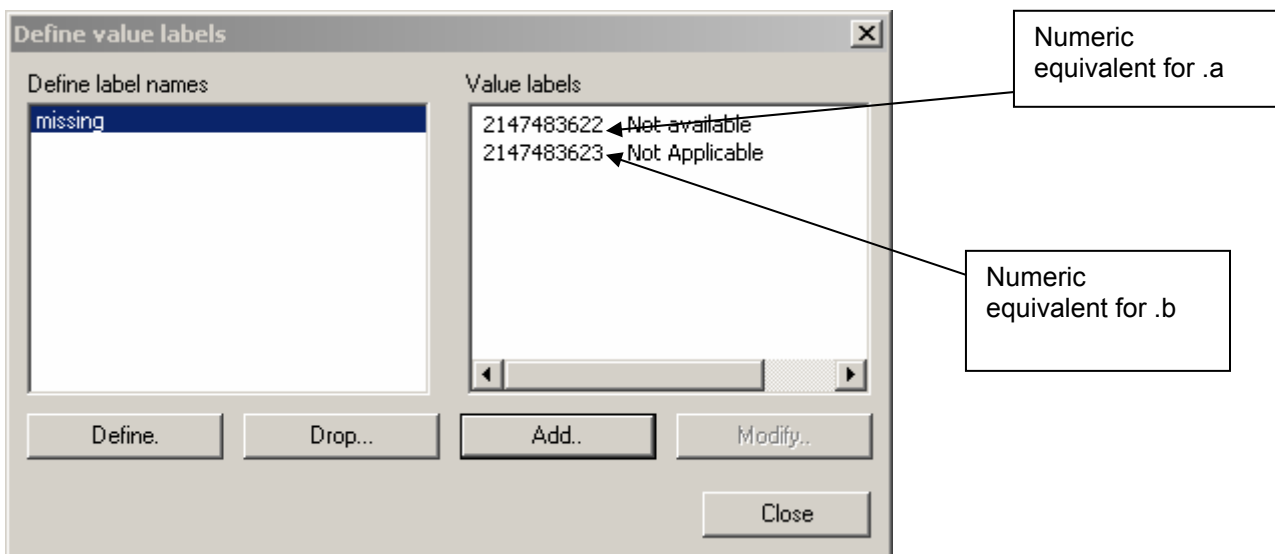
Note that the system missing is '.'. This is acceptable if you only have 1 level of missing data, but sometimes we want to differentiate between not applicable, not available, etc.

Open the **STATA** Data Editor again, and this time, write '.a' instead of '.'.

```
. list, separator(6)
```

	id	age	sex	var1	var2	var3
1.	1	22	M	4	2	1
2.	2	40	F	2	3	1
3.	3	27	M	3	3	2
4.	4	35	M	2	2	4
5.	5	24	F	1	2	2
6.	6	.a	F	1	2	3

You now can create a value label for the missing data and attach it to the **age** variable.



Remember to attach the label to the variable. This is the result of the labelling.

```
. list, separator(6)
```

	id	age	sex	var1	var2	var3
1.	1	22	M	4	2	1
2.	2	40	F	2	3	1
3.	3	27	M	3	3	2
4.	4	35	M	2	2	4
5.	5	24	F	1	2	2
6.	6	Not available	F	1	2	3

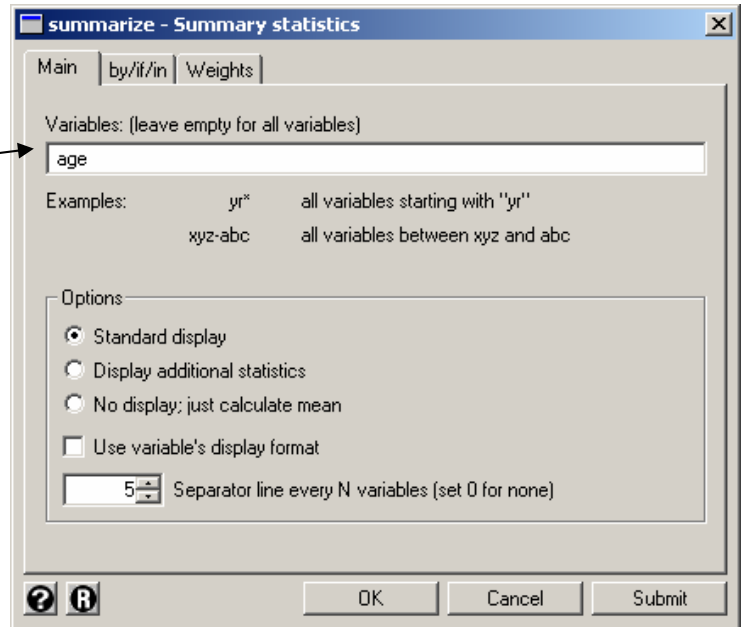
STATA will automatically exclude any missing data from the analysis. So for example, click on

Statistics > **Summaries, tables & tests** > **Summary statistics** > **Summary statistics**

to find the mean.

Choose *age*.

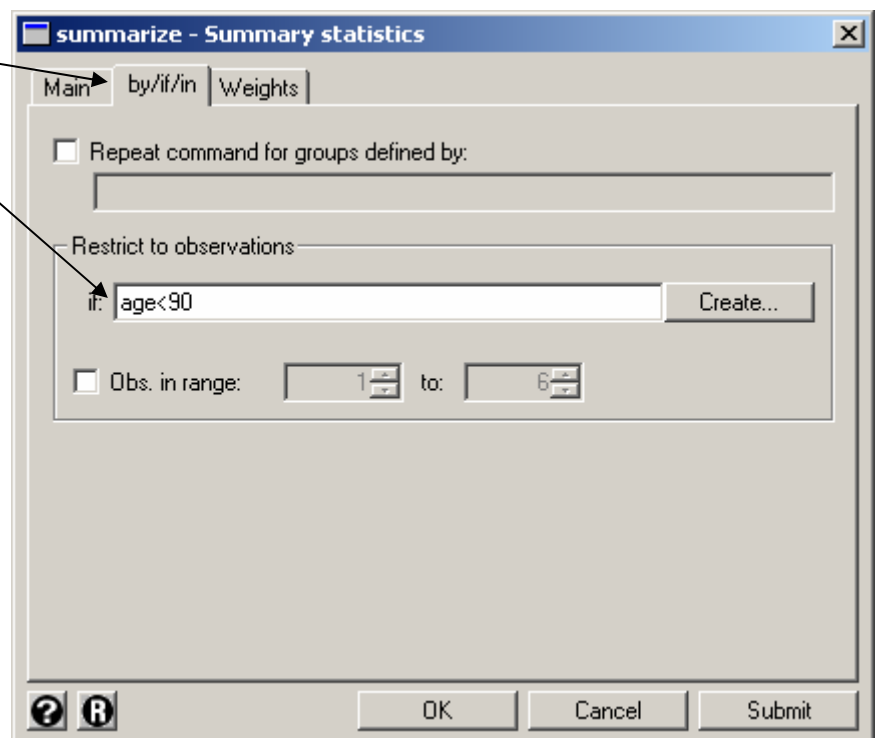
5 observations, excluding missing value



```
. summarize age
```

Variable	Nbs	Mean	Std. Dev.	Min	Max
age	5	29.6	7.635444	22	40

If we had chosen all ages less than 90, to manually eliminate the missing data



We would have obtained the same answer again.

```
. summarize age if age<90
```

Variable	Obs	Mean	Std. Dev.	Min	Max
age	5	29.6	7.635444	22	40

Crosstabulation in STATA

The crosstabulation procedure allows you to explore the relationship between, normally just two, categorical variables. It will show you a table of the joint frequency distributions of the two variables. Accompanying statistics will also tell you if there is a significant association between the two variables.

As an example of crosstabulation, we will use the file 'sample.dta' which you should have created in the last practical session. Make sure that the missing codes have been labeled as '.a' or '.b'.

We will crosstabulate the two variables **hincdiff** (How well are you managing your income?) with **srinc** (Which income group would you place yourself?).

In order to carry out the cross tabulation, click on

Statistics > Summaries, tables & tests > Tables > Table of summary statistics (table)

The screenshot shows the 'Table of summary statistics' dialog box in STATA. The 'Row variable' is set to 'hincdiff' and the 'Column variable' is set to 'srinc'. The 'Statistics' section has five rows, each with a 'None' dropdown. The 'Percentile' section has five rows, each with a '50' dropdown. The 'Variable' section has five empty text boxes. The 'OK' button is highlighted. Three callout boxes provide instructions: 'hincdiff was selected as the row variable.', 'srinc was selected as the column variable.', and 'Press OK.'

As a rule of thumb, you should place the **dependent** variable as the row variable and the **independent** variable as the column variable. In this example

it is assumed, if anything, that it is high income or lack of it which affects how people feel about whether they are managing, not that how they feel they are managing affects their income.

```
. table hincdiff srinc
```

hincdiff	srinc	
	Middle Income	Low Income
Very well	1	?
Quite well	10	3
Not very well	1	2
Not at all well		

STATA ignores all missing values

This indicates that no person with low income is managing very well.

This command only displays the cross tabulation between the two variables. In most of the cases, we will also be interested in percentages as well as measures of association. Click on

Statistics > Summaries, tables & tests > Tables > Two-way tables with measures of association

Choose the row and column variable

Click for column percentages

Click OK

After that the resulting table looks like:

```
. tabulate hincdiff srinc, column
```

Key			
	frequency	column	percentage
hincdiff	srinc		Total
	Middle In	Low Incom	
Very well	1 8.33	0 0.00	1 4.17
Quite well	10 83.33	7 58.33	17 70.83
Not very well	1 8.33	3 25.00	4 16.67
Not at all well	0 0.00	2 16.67	2 8.33
Total	12 100.00	12 100.00	24 100.00

We now can say that about 83% of those who said they were on a middle income are managing quite well.

Recoding

When we look at the table, we notice that it has two empty cells. A reasonable option to decrease the number of empty cells would be to collapse across some categories of the variable **hincdiff**, i.e. 'Very well' and 'Quite well' could be collapsed into one category called 'Well'. While 'Not very well' and 'Not at all well' could also be collapsed into one 'Not well' category. For this we use the recoding facility of **STATA**.

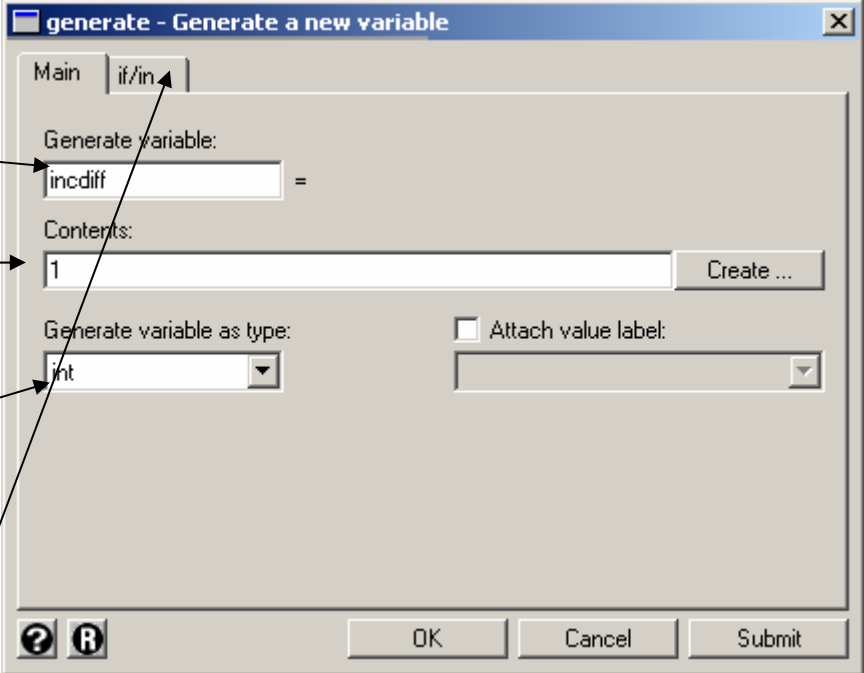
There are other reasons to recode such as:

- Altering an existing coding scheme, e.g. to regroup a continuous variable like age
- During editing to correct coding errors, e.g. to change any wild (i.e. erroneous) codes to a missing value

When recoding, it is always advisable to create a new variable so that if any errors occur while recoding, you can still go back to your original variable and re-start recoding.

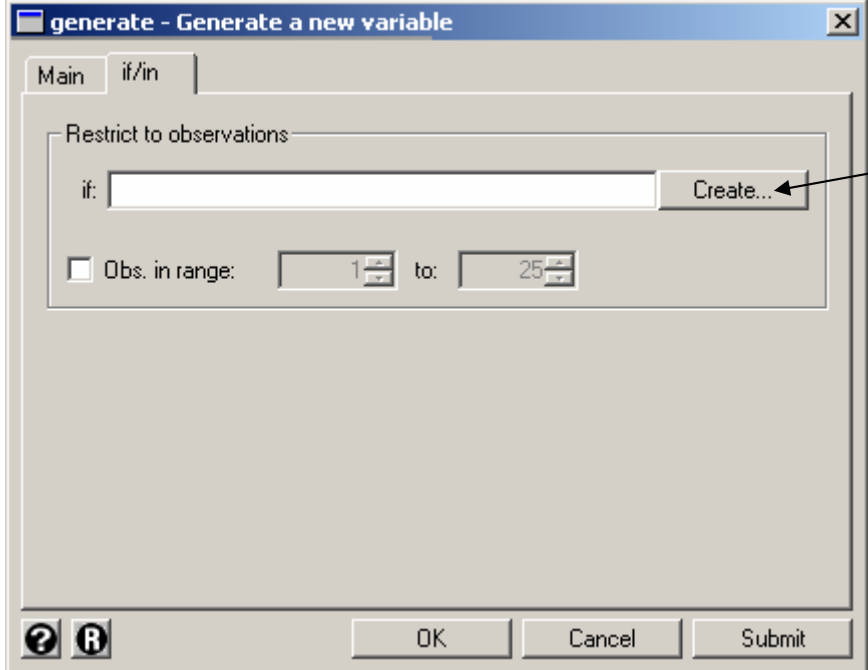
To illustrate recoding, click on

Data > Create or change variables > Create new variables



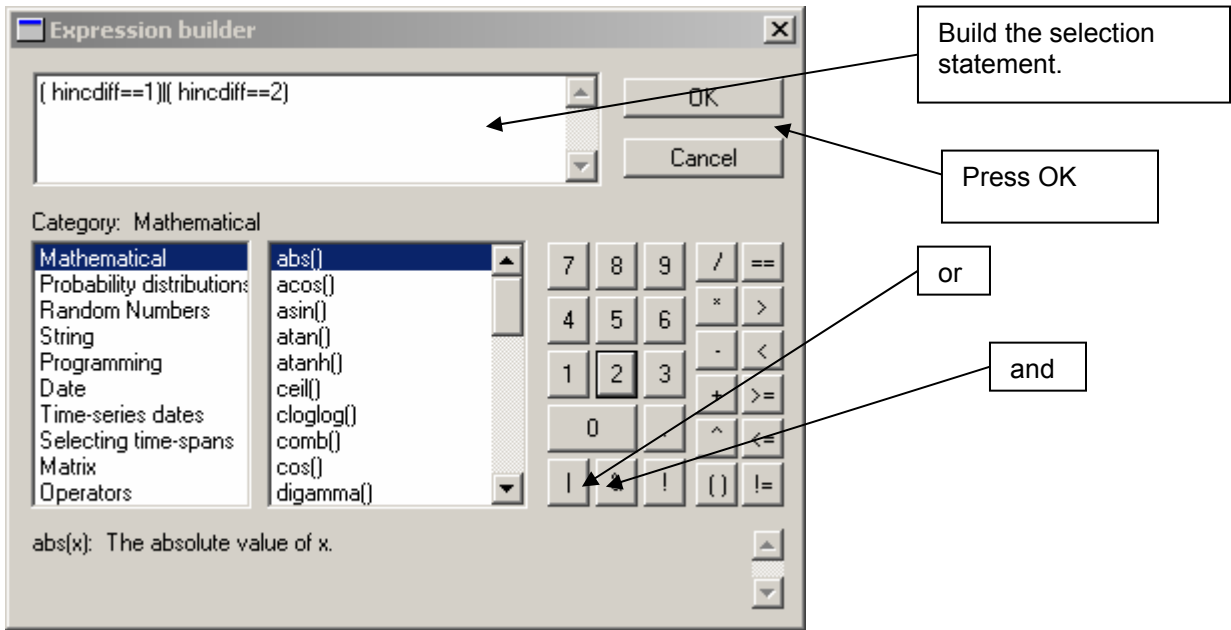
The screenshot shows the 'generate - Generate a new variable' dialog box. The 'Main' tab is selected, and the 'if/in' sub-tab is active. The 'Generate variable:' field contains 'incdiff'. The 'Contents:' field contains '1'. The 'Generate variable as type:' dropdown is set to 'int'. The 'Attach value label:' checkbox is unchecked. Annotations with arrows point to the following elements:

- Name the new variable (points to 'incdiff')
- Type the 1st value of the recode. (points to '1')
- Change to indicate that **incdiff** is an integer. (points to the 'int' dropdown)
- Click on **if/in**. (points to the 'if/in' sub-tab)



The screenshot shows the 'generate - Generate a new variable' dialog box with the 'Restrict to observations' section expanded. The 'if:' field is empty, and the 'Create...' button is highlighted. The 'Obs. in range:' checkbox is unchecked, and the range is set from 1 to 25. An annotation with an arrow points to the 'Create...' button:

- Click on **Create** to select the cases which will be recoded.

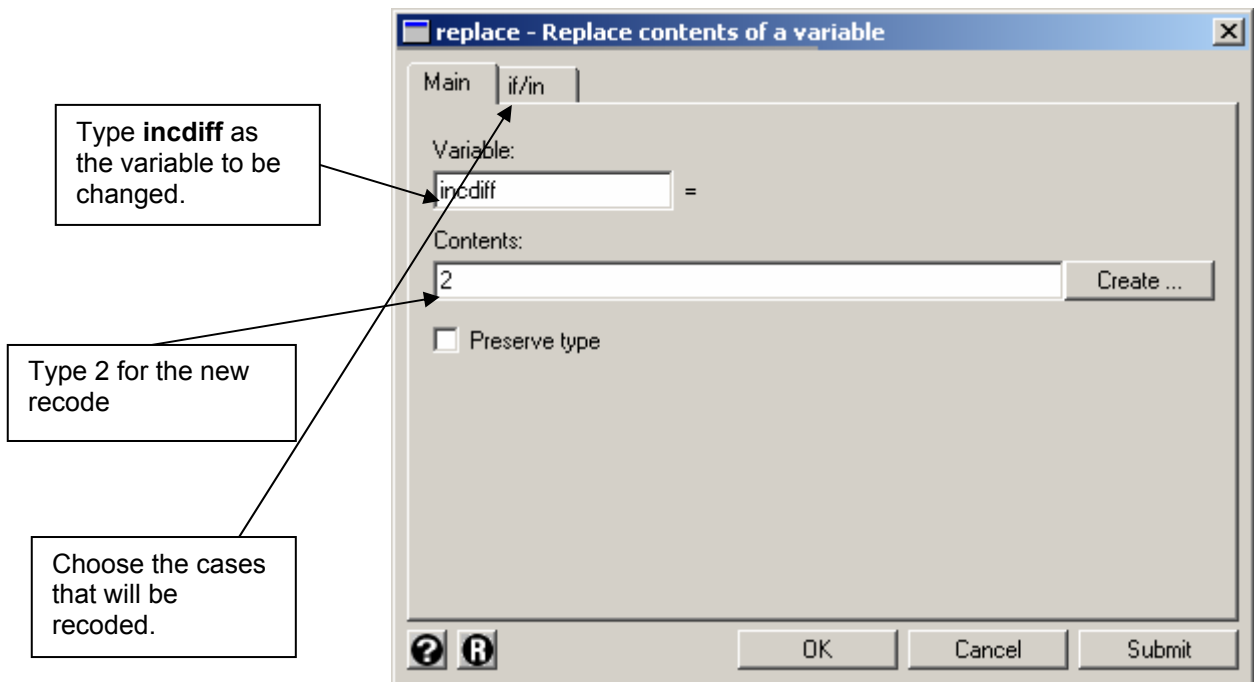


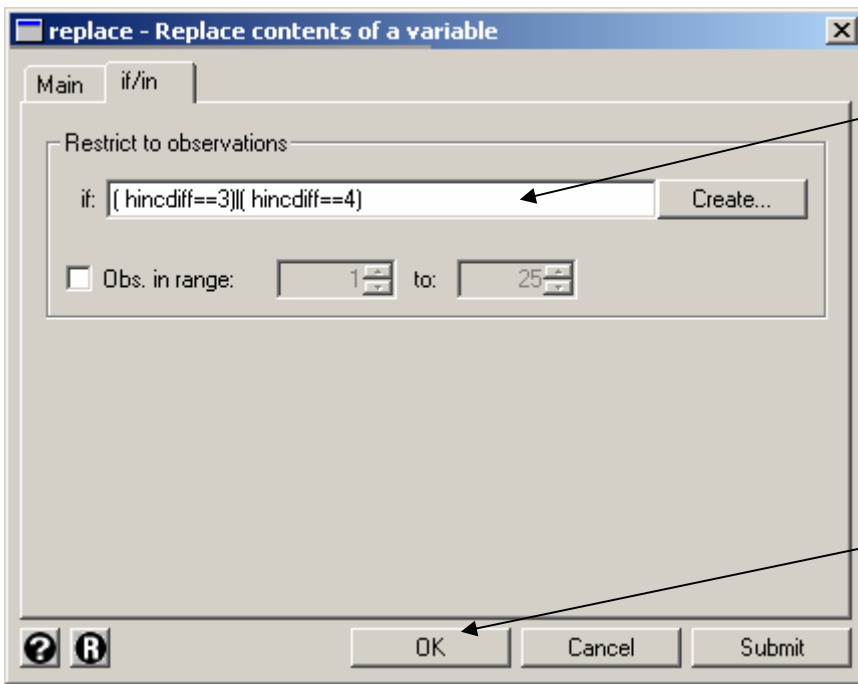
The Results window indicates whether the new variable has been created:

```
. generate int incdiff= 1 if < hincdiff==1>!< hincdiff==2>
<? missing values generated>
```

This is not ready yet as only 1 recode has been done. We need now to recode values 3 and 4 into 2. The variable *incdiff* now exists, and therefore click on

Data > Create or change variables > Change contents of variable





Use **create** to select the cases that will be recoded.

Click **OK**

```
. replace incdiff = 2 if < hincdiff==3)!< hincdiff==4)
<6 real changes made>
```

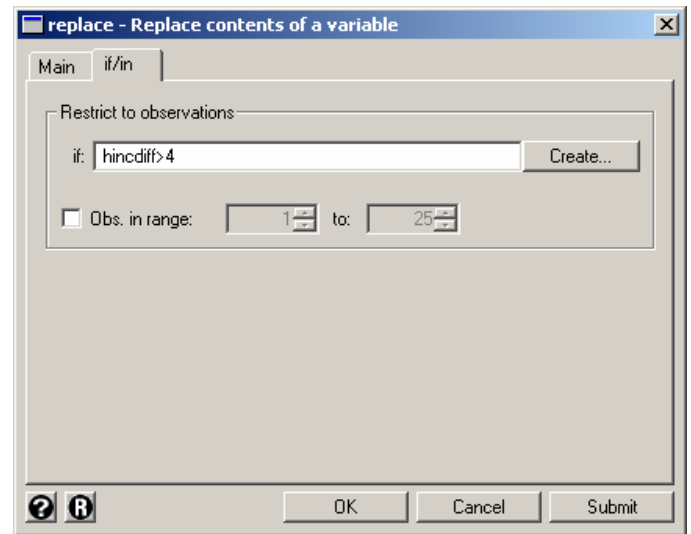
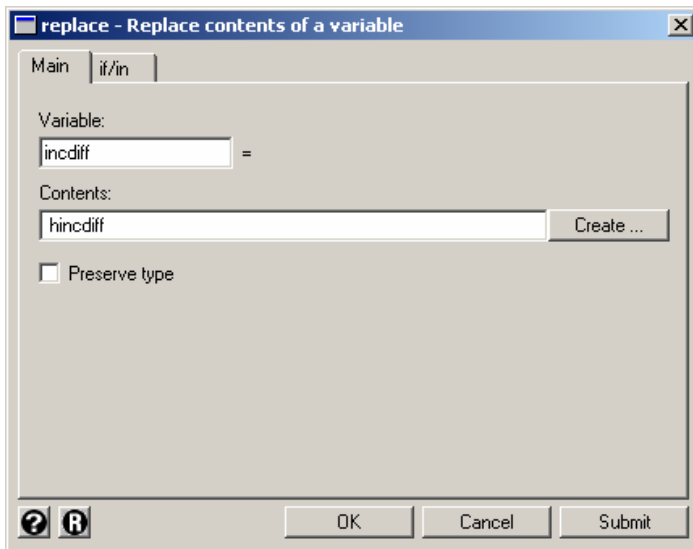
```
. list hincdiff incdiff, separator(0)
```

	hincdiff	incdiff
1.	Quite well	1
2.	Quite well	1
3.	Not at all well	2
4.	Quite well	1
5.	Not very well	2
6.	No response	.
7.	Not very well	2
8.	Quite well	1
9.	Quite well	1
10.	Not at all well	2
11.	Quite well	1
12.	Not very well	2
13.	Quite well	1
14.	Quite well	1
15.	Quite well	1
16.	Quite well	1
17.	Quite well	1
18.	Quite well	1
19.	Quite well	1
20.	Quite well	1
21.	Not very well	2
22.	Quite well	1
23.	Very well	1
24.	Quite well	1
25.	Quite well	1

Values 1 and 2 recoded to 1, values 3 and 4 recoded to 2

Missing data not recoded

We need to re-run the procedure for all the missing data:



Also note that the value labels need now to be changed. Therefore, after copying the missing values and creating new labels, the new variable should look like this.

```
list hincdiff incdiff, separator(0)
```

	hincdiff	incdiff
1.	Quite well	Well
2.	Quite well	Well
3.	Not at all well	Not well
4.	Quite well	Well
5.	Not very well	Not well
6.	No response	No response
7.	Not very well	Not well
8.	Quite well	Well
9.	Quite well	Well
10.	Not at all well	Not well
11.	Quite well	Well
12.	Not very well	Not well
13.	Quite well	Well
14.	Quite well	Well
15.	Quite well	Well
16.	Quite well	Well
17.	Quite well	Well
18.	Quite well	Well
19.	Quite well	Well
20.	Quite well	Well
21.	Not very well	Not well
22.	Quite well	Well
23.	Very well	Well
24.	Quite well	Well
25.	Quite well	Well

If we now repeat the crosstabulation, but we choose *incdiff* rather than *hincdiff* as the row variable, we would obtain the following table.

```
. tabulate incdiff srinc, column
```

Key		Frequency		column	percentage
incdiff	srinc		Total		
	Middle In	Low Incom			
Well	11 91.67	7 58.33	18 75.00		
Not well	1 8.33	5 41.67	6 25.00		
Total	12 100.00	12 100.00	24 100.00		

You can see that we have removed the empty cells from the cross tabulation.

Another way to Recode

In **STATA** we can recode to the same variable, rather than creating a new variable. Look again at the data file 'sample.dta'. Click on

Data > **Create or change variables** > **Other variable transformation commands** > **Recode categorical variables**

Enter the categorical variable that you will recode.

Click to obtain the rules for recoding.

The rules for recoding are given in the following table:

<i>rule</i>	Example	Meaning
(# = #)	(3 = 1)	3 recoded to 1
(# # = #)	(2 . = 9)	2 and . recoded to 9
(## = #)	(1/5 = 4)	1 through 5 recoded to 4
(nonmissing = #)	(nonmiss = 8)	all other nonmissing to 8
(missing = #)	(miss = 9)	all other missings to 9

Therefore, use the rules to recode 1 and 2 to 1 and 3 and 4 to 2.

The screenshot shows the 'recode - Recode categorical variable' dialog box with 'hincdiff' selected. The 'Required' rule is '(1 2 = 1)' and the 'Optional' rule is '(3 4 = 2)'. Below the dialog, a terminal window displays the command `. list hincdiff, separator<0>` and the resulting output for 25 cases. The output shows the recoded values for 'hincdiff'.

Case	hincdiff
1.	Very well
2.	Very well
3.	Quite well
4.	Very well
5.	Quite well
6.	No response
7.	Quite well
8.	Very well
9.	Very well
10.	Quite well
11.	Very well
12.	Quite well
13.	Very well
14.	Very well
15.	Very well
16.	Very well
17.	Very well
18.	Very well
19.	Very well
20.	Very well
21.	Quite well
22.	Very well
23.	Very well
24.	Very well
25.	Very well

Click on OK to submit the change. Note that ***hincdiff*** will now change to the new variable. Note also that you have to change the value labels, as these still reflect the old ***hincdiff***.

The only way to get ***hincdiff*** back is to reload the data. So it might be wise to 1st copy ***hincdiff*** to ***incdiff***, and then modify ***incdiff***.

Computing New Variables

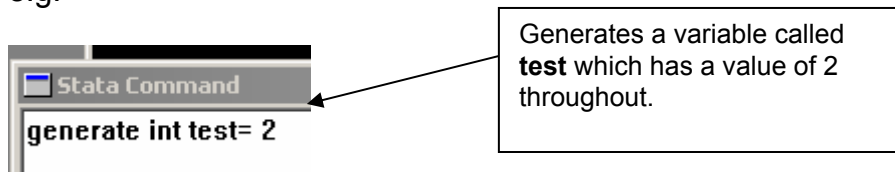
Using the '**generate**' and '**replace**' commands, we can create new variables and assign values to these variables for each case.

The basic command would be

Basic Command

generate type New variable = mathematical expression

e.g.



Before we look at further examples, let us take a look at the types of mathematical expressions that we might have.

The mathematical expressions can be...

- **A variable**

AGEGROUP = AGE

This allows you to create a copy of another variable.

- **A constant**

TOTINC = 0

This may be useful if you want to set a variable to 0, such as TOTINC (total income) before you then go on and use a more complicated command to calculate the actual total income.

A mathematical expression can include an arithmetic operator

- + addition
- subtraction
- * multiplication
- / division
- ^ exponentiation (to the power of)

Some examples

TOTINC = WAGES + BONUS

YEARS = MONTHS/12

SQDOCTOR=DOCTOR^2

BYEAR = 87 - RAGE

In the last example, we can discover the birth year of the respondents in the 1987 Social Attitude Survey, knowing their age (RAGE).

- **Arithmetic Functions**

i.e. LG10 or SQRT

LGINCOME = LG10(TOTINC)

Will calculate the log of the variable TOTINC and put the value into the new variable LGINCOME.

- **Matrix Functions**

trace(A)

will calculate the sum of the diagonal elements of matrix A.

Example

The file '*wages.dta*' contains information on 4 hypothetical people. For each respondent we have the income they earned and the bonus payments.

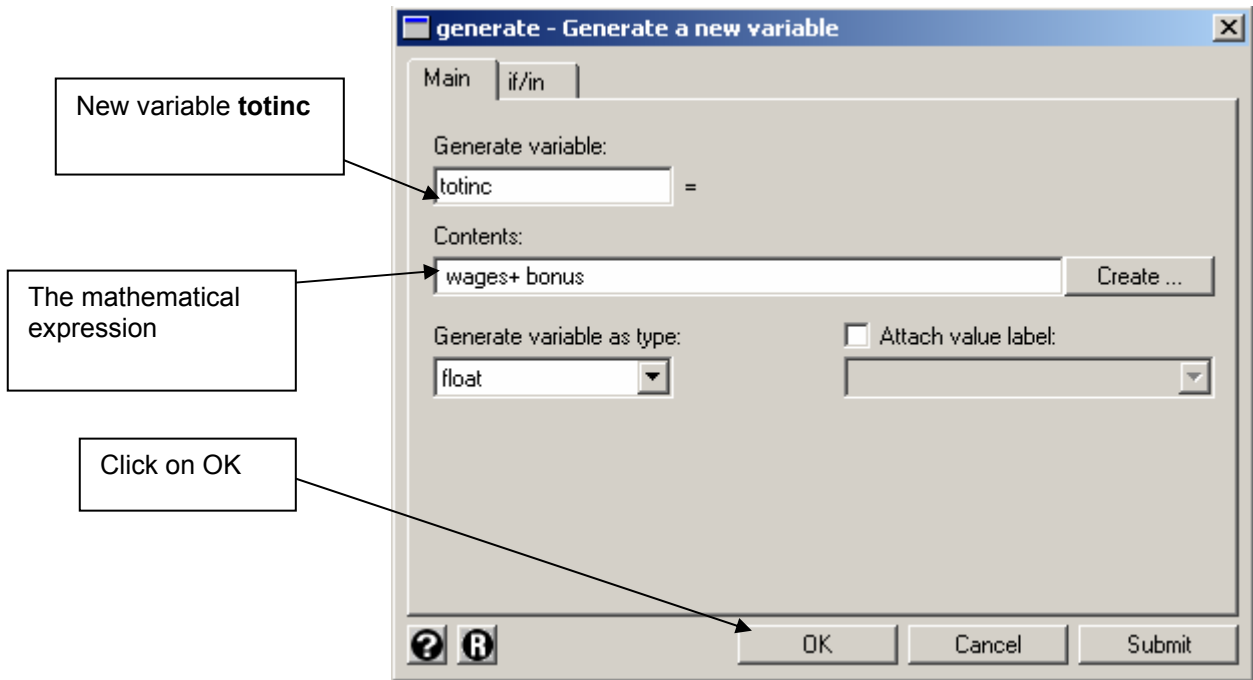
```
. list, separator(5)
```

	wages	bonus
1.	12000	234
2.	9500	0
3.	17500	350
4.	1575	120

Suppose we wish to create a new variable, called '*totinc*', which will be the sum of *wages* and *bonus*.

In **STATA** we click on

Data > **Create or change variables** > **Create new variable**



```

. generate float totinc= wages+ bonus
. list, separator(5)

```

	wages	bonus	totinc
1.	12000	234	12234
2.	9500	0	9500
3.	17500	350	17850
4.	1575	120	1695

List the variables to show that **totinc** has been created correctly.

Selecting Cases

We might sometime which to perform an analysis on a subset of cases, e.g. only women or only married people. Let us open the data set 'bsas91.dta'.

The method for selecting cases will be similar to the method used before to recode. Some simple conditions are:

rsex = 2

to choose all the female respondents

rsex = 2 & marstat = 2

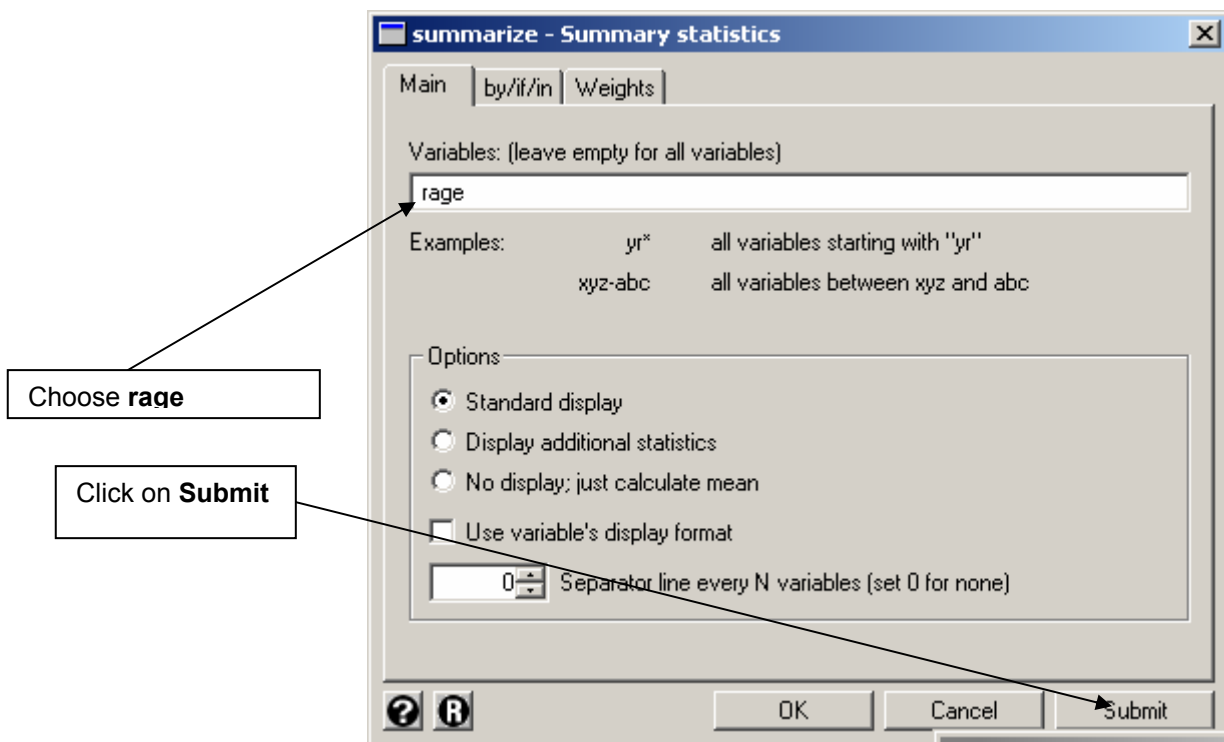
to choose all the female respondents who are living as married

prsoccl < srsoccl

to choose the respondents where the parents social class is less than respondents social class which because of the way class is coded (1 is high 6 is low) means those cases where downward social mobility has occurred.)

Let us obtain the average age of the respondents. This is done by clicking on

Statistics > Summaries, Tables & Tests > Summary Statistics > Summary Statistics



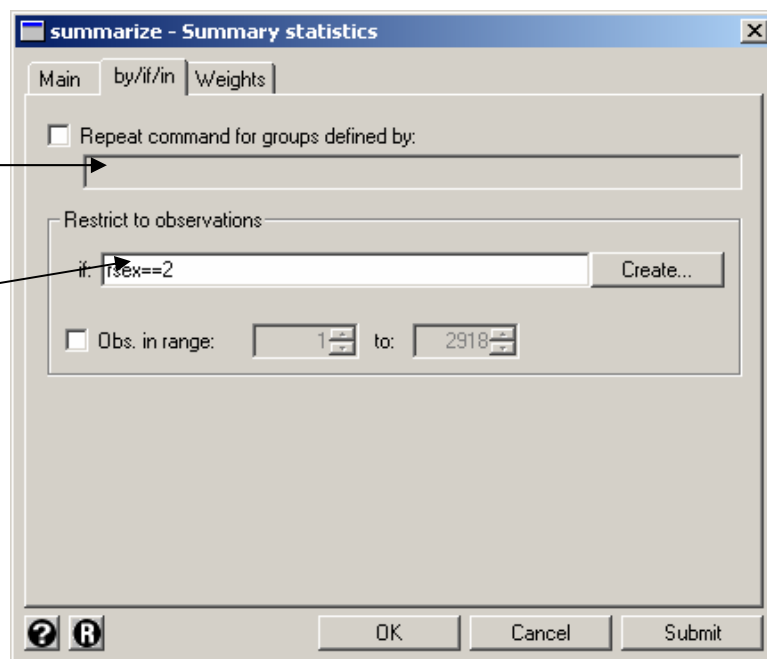
```
. summarize rage, separator(0)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
rage	2905	47.73219	18.26468	18	94

This indicates that the mean for all the dataset (excluding those missing) is 47.73219. If we wanted to check whether the mean is higher or lower for females, then click on the **by/if/in** tab.

Use if you wanted to split the display.

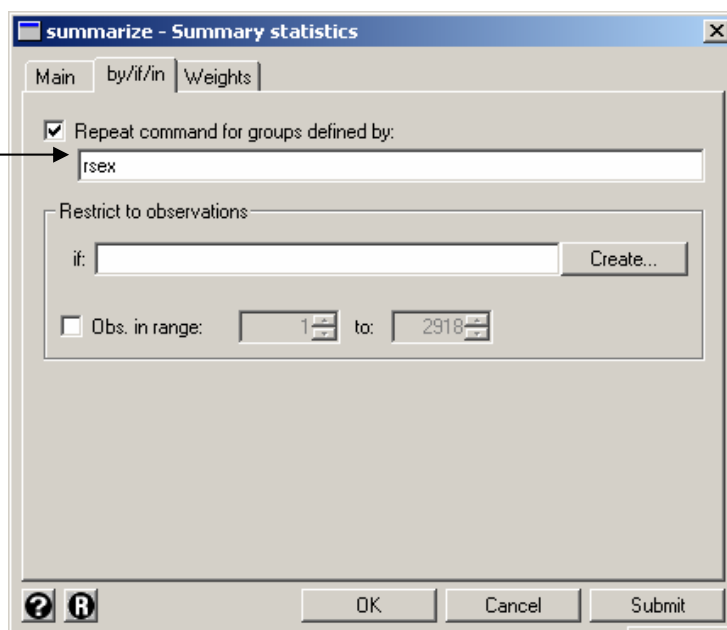
Choose only where rsex is 2 (or female)



```
. summarize rage if rsex==2, separator(0)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
rage	1617	47.91775	18.42946	18	94

Instead of filtering for females, we could obtain separate output for females and males.



```
>
-> rsex = 1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
rage	1288	47.49922	18.06012	18	92

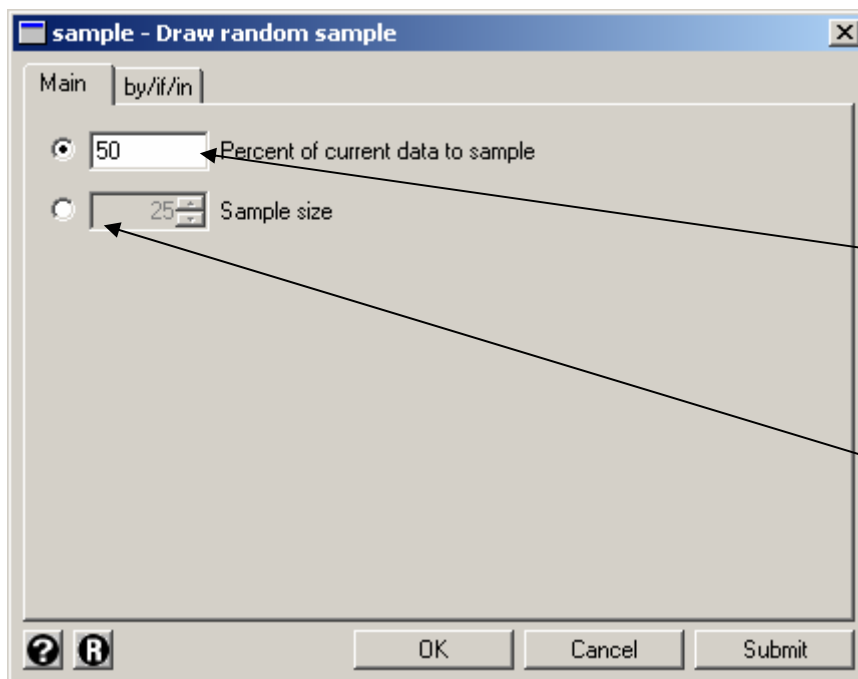
```
>
-> rsex = 2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
rage	1617	47.91775	18.42946	18	94

Sampling Cases

If you were working with a very large data set it might be advisable to try out your analysis on a sample before using the whole data set. This can be an enormous saving in processing time. To sample cases, click on

Statistics > Resampling & simulation > Draw a random sample



Choose **50%** of the current data in the sample.

You can also choose an exact number.

```

. summarize rage

```

Variable	Obs	Mean	Std. Dev.	Min	Max
rage	2905	47.73219	18.26468	18	94

```

. sample 50
(1459 observations deleted)
. summarize rage

```

Variable	Obs	Mean	Std. Dev.	Min	Max
rage	1451	47.84493	18.05471	18	94

This shows that the number of observations before sampling was done was 2905, then after it decreased to 1451. The original file has been lost so it is advisable to first make a copy of the file before attempting to sample.

Split Analysis

STATA has the facility to enable you to split your data file into separate groups for analysis. For instance, if the file was split according to the variable **rrgclass**, respondents social class according to the Registrar Generals Classification, and then you asked for the frequencies of the variable **rsex**, you would end up with a frequency table for each social class. Using the split command is equivalent to separately selecting each category of social class and then running the frequencies command.

Alternatively, you may wish to perform a particular analysis based not only on the sex of the respondent but also on their age, say, whether they are above or below 40. In other words, you want to split your file based on **two** variables. Suppose you wanted a separate frequency tables for the following subgroups in '**bsas91.dta**'.

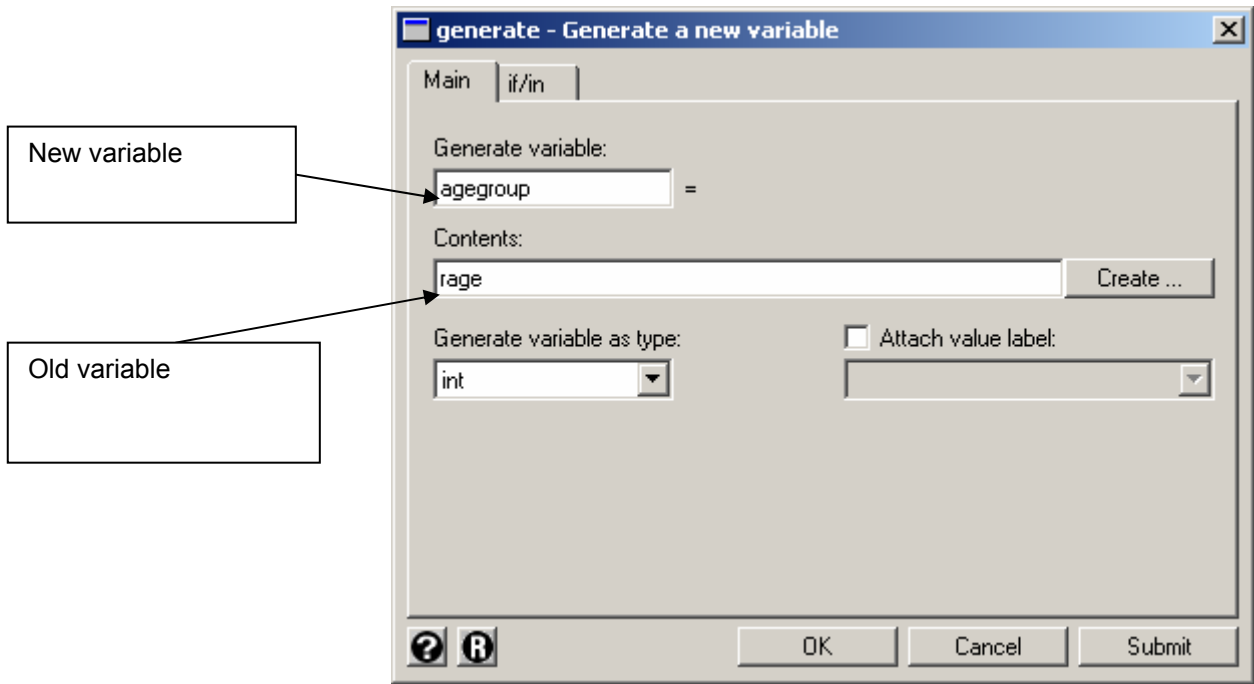
males under 40
 males 40 or over
 females under 40
 females 40 or over

First we need to recode the variable **rage** into two categories, below 40 and equal to or above 40. Let us call this new variable, **agegroup**.

Click on

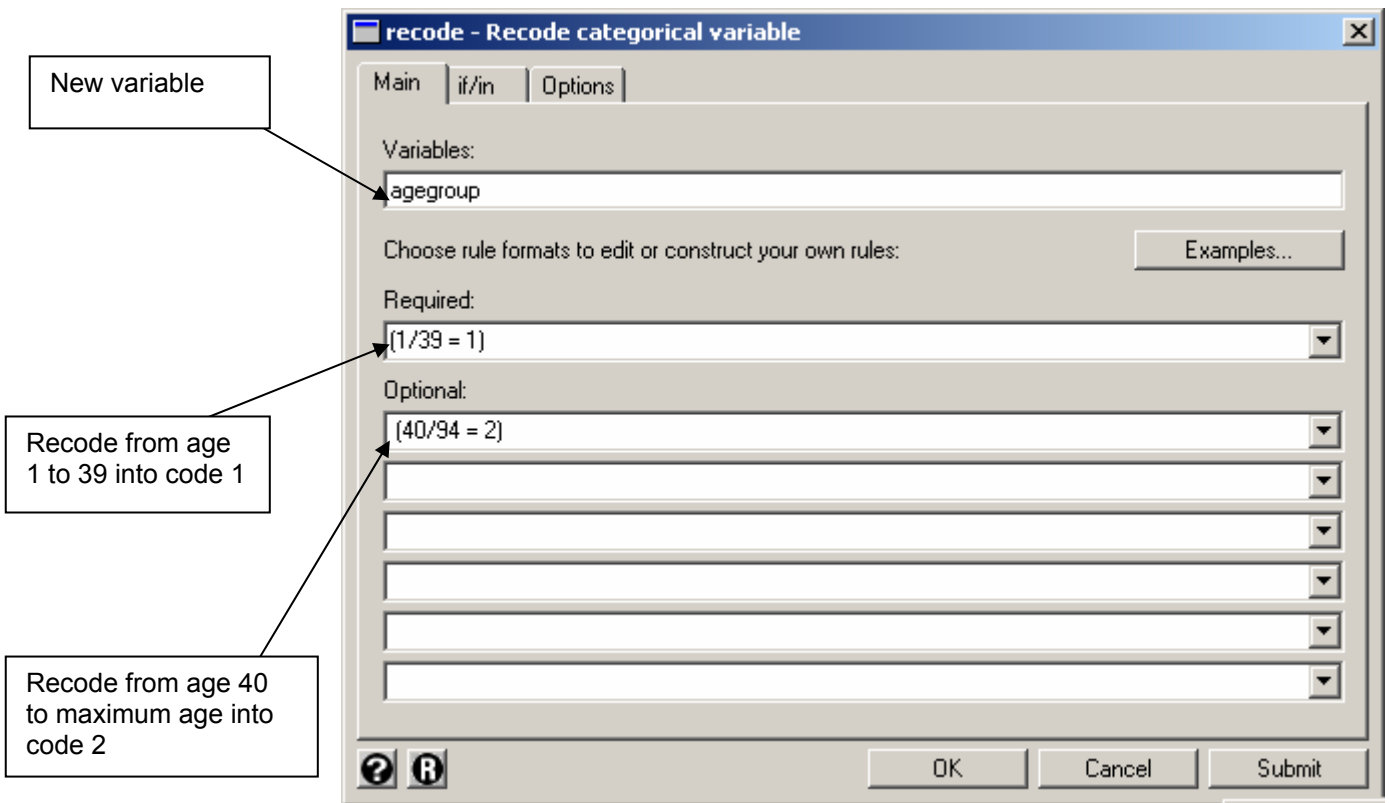
Data > **Create or change variables** > **Create new variable**

to create a copy of **rage** into **agegroup** as seen in the following window.



Then click on

Data > **Create or change variables** > **Other variable transformation commands** > **Recode categorical variables**



If we obtain a frequency distribution of the new variable, we have

```
. table agegroup
```

agegroup	Freq.
1	1,105
2	1,800

Suppose we want now to obtain a frequency distribution of the variable *srinc* (income group) for each of the 4 groups.

Choose the variable you want the frequency of.

Click to enter the split details.

Choose **agegroup** and **rsex** as the split variables.

This is the output obtained:

```
. hysort agegroup rsex: table srinc
```

```
-> agegroup = 1, rsex = 1
```

SRINC	Freq.
1	4
2	135
3	87
8	1
9	3

```
-> agegroup = 1, rsex = 2
```

SRINC	Freq.
1	8
2	148
3	165
8	2
9	2

```
-> agegroup = 2, rsex = 1
```

SRINC	Freq.
1	16
2	178
3	192
9	4

```
-> agegroup = 2, rsex = 2
```

SRINC	Freq.
1	12
2	205
3	264
8	6
9	4

```
-> agegroup = .a, rsex = 1
```

no observations

```
-> agegroup = .b, rsex = 1
```

SRINC	Freq.
3	3

```
-> agegroup = .b, rsex = 2
```

SRINC	Freq.
1	1
2	2
3	3

Note that we obtained output even for the missing data.

Practical Session 2

1. Income and perception of living standards

In this exercise, you will start by re-running the table, but this time using a subset of the 1991 data set containing data for 2836 respondents. Then you will be using one of the **STATA** data transformation commands to 'recode' some of the variables from that dataset.

Load the file '*bsas91.dta*'. Crosstabulate **hincdiff** with **srinc**. Recode **hincdiff** into **incdiff** as in the example of page 2-7.

Add appropriate value labels to **incdiff**.

Crosstabulate **incdiff** with **srinc**. Obtain **Column Percentages** for this crosstabulation.

Is it the case that richer respondents are likely to think that they are coping better than poor ones? While you should have had some idea about an answer to this question from the tiny sample used previously, you should now be able to answer the question with some confidence.

2. Political identification and age

The variable **partyid1** records the political identification of the respondent (note that the variable is spelt with ID (the letters I and D) and a final digit, 1). The variable shows respondents' answers to the question:

What political party do you support, or feel a little closer to, or if there was a general election tomorrow, which one would you most likely support?

How does party identification vary with age? Carry out the following steps:

Remove the 4 levels of missing data in the variable. Refer to the code book supplied as an appendix to the notes.

Obtain a frequency distribution of the variable **partyid1** to see the range of parties and the distribution of respondents between them.

Recode all those who identify with the Scottish Nationalists, Plaid Cymru, Other Parties, and who gave Other Answers or No answer into the missing category (code 9). Call the new variable **polpart**, i.e. Recode **partyid1** ≥ 6 to 9, and copy everything else as is.

Recode **rage** into a different variable **agegp** by dichotomizing it into 2 groups; those aged 40 or over and those under 40, you will need to decide what to do with **No response**, coded 99.

Add appropriate value labels to *polpart* and *agegp*. Remember to indicate the missing data.

Crosstabulate political identification (*polpart*) with age group (*agegp*).

Are older respondents more likely to vote Conservative than younger ones? Where was the Alliance support concentrated?

Save your data set as '*newbsas.dta*'. Do not change '*bsas91.dta*'.

3. Political identification and age

Use the file '*bsas91.dta*'. The British Social Attitudes Survey includes a set of variables about respondents' opinion about the seriousness of various environmental pollutants and damage (noise from aircraft, lead from petrol, industrial waste in rivers and seas, waste from nuclear electricity stations, industrial fumes in the air, noise and dirt from traffic, acid rain, aerosol chemicals and loss of rain forests). Respondents were asked to indicate, for each of these whether they thought the effect on the environment was not at all serious (code 1), not very serious (code 2), quite serious (code 3), very serious (code 4) or that they did not know (code 8) or did not reply (code 0). The answers are recorded in variables called *envir1*, to *envir9*.

One way of getting an overall, summary score for a respondent's attitude to the environment would be to sum the scores on these seven variables. This can be done with the generate command in which a new variable, *envirall* is set to the sum total of the scores on each of the *envir* variables, for each respondent.

Be careful that the *envir* variables are not coded as string variables. If this is the case, then a normal summation on string is not the same as an addition of numbers. You might want to change the string variables to numeric variables by clicking on

Data ➤ Create or change variables ➤ Other variable transformation commands ➤ Convert variables from string to numeric

Now you can create a new numeric variable *envirall* by

***envirall = envir1 + envir2 + envir3 + envir4 + envir5 + envir6 + envir7 +
envir8 + envir9***

4. Mobility tables

An 'inter-generational social mobility' table cross-tabulates parents' class by respondents' class, to show the extent to which a society is open or closed to movement through the class structure. Most mobility tables studied in the research literature have examined fathers' class against sons' class and have ignored the class of mothers and daughters. This is partly because women have for so long been almost ignored by sociologists, but also because class is normally assessed on the basis of respondents' occupation and until the 1960s the majority of women were not in paid employment.

Usually mobility tables are constructed from data about people's actual occupations categorized into social classes. In the BSAS dataset, however, the only data on parents' social class comes from respondents' *own rating* of their parents' social class. In some ways this is less satisfactory than occupational data (the ratings may well be confounded by the respondents' own positions in the class structure, for instance), but one of the requirements of secondary analysis of data collected by other people is that one has to make the best of what one has got.

A complication with the interpretation of mobility tables is that the occupational and class structure has changed significantly over the course of the century. In a representative sample of the population, there will be some young respondents whose fathers are still alive and working, and some old respondents whose fathers retired near the beginning of the century from an occupational structure very different from the present one. Thus a variable about the social class of fathers will be a rather messy composite, holding some data about fathers whose class is assessed in terms of a class structure which no longer exists and some data about fathers whose class is assessed in terms of the present structure. One tactic for getting over this problem is to include only respondents within a particular age range.

Open the data set '*bsas91.dta*'. In each analysis we have to select only those aged between 18 and 40.

Obtain a crosstabulation of parent's social class (**prsoccl**) by own social class (**srsoccl**).

What percentage of respondents with working class parents now think of themselves as middle class?

The table you have just obtained includes both male and female respondents. However, the class structure and the mobility of men and women are very different. It would make more sense to look at separate mobility tables for the two sexes.

Click on

Statistics > **Summaries, tables & tests** > **Tables** > **All possible two-way tabulations**

Choose *prsoccl* and *srsoccl* and the 2 variables

Click on this tab

tab2 - Two-way tables

Main by/if/in Weights Advanced

Categorical variable(s):
prsoccl srsoccl

Test statistics

- Pearson's chi-squared
- Fisher's exact test
- Goodman and Kruskal's gamma
- Likelihood-ratio chi-squared
- Kendall's tau-b
- Cramer's V

Cell contents

- Pearson's chi-squared
- Within-column relative frequencies
- Within-row relative frequencies
- Likelihood-ratio chi-squared
- Relative frequencies
- Expected frequencies
- Suppress frequencies

Treat missing values like other values

Do not wrap wide tables

Suppress the cell contents key

Suppress displaying the value labels

OK Cancel Submit

Choose *rsex* as the grouping variable

tab2 - Two-way tables

Main by/if/in Weights Advanced

Repeat command for groups defined by:
rsex

Restrict to observations

if: Create...

Obs. in range: to:

OK Cancel Submit

Compare the resulting tables for men and for women. Is upward mobility more or less likely for men than for women?

What reservations is it necessary to make about drawing conclusions from these data?